

EXHIBIT 1

(Part 2 of 3)

US 6,355,432 B1

1

PRODUCTS FOR DETECTING NUCLEIC ACIDS

CROSS REFERENCE TO RELATED APPLICATIONS

This is a continuation of application Ser. No. 09/362,089, filed Jul. 28, 1998, pending; which is a divisional of application Ser. No. 09/056,927, filed Apr. 8, 1998, now U.S. Pat. No. 6,197,506; which is a continuation of application Ser. No. 08/670,118, filed Jun. 25, 1996, now U.S. Pat. No. 5,800,992; which is a divisional of application Ser. No. 08/168,904, filed Dec. 15, 1993, now abandoned; which is a continuation of application Ser. No. 07/624,114, filed Dec. 6, 1990, now abandoned; each of which is hereby incorporated by reference.

Additional commonly assigned application No. 07/492,462, filed Mar. 7, 1990, now U.S. Pat. No. 5,143,854; application No. 07/362,901, filed Jun. 7, 1989, now abandoned; application Ser. Nos. 07/624,120 and 07/626,730, both of which were filed on Dec. 6, 1990; application Ser. No. 07/435,316, filed Nov. 13, 1989, now abandoned; and U.S. Pat. No. 5,252,743 are also hereby incorporated herein by reference.

BACKGROUND OF THE INVENTION

The present invention relates to the sequencing, fingerprinting, and mapping of polymers, particularly biological polymers. The inventions may be applied, for example, in the sequencing, fingerprinting, or mapping of nucleic acids, polypeptides, oligosaccharides, and synthetic polymers.

The relationship between structure and function of macromolecules is of fundamental importance in the understanding of biological systems. These relationships are important to understanding, for example, the functions of enzymes, structural proteins, and signalling proteins, ways in which cells communicate with each other, as well as mechanisms of cellular control and metabolic feedback.

Genetic information is critical in continuation of life processes. Life is substantially informationally based and its genetic content controls the growth and reproduction of the organism and its complements. Polypeptides, which are critical features of all living systems, are encoded by the genetic material of the cell. In particular, the properties of enzymes, functional proteins, and structural proteins are determined by the sequence of amino acids which make them up. As structure and function are integrally related, many biological functions may be explained by elucidating the underlying structural features which provide those functions. For this reason, it has become very important to determine the genetic sequences of nucleotides which encode the enzymes, structural proteins, and other effectors of biological functions. In addition to segments of nucleotides which encode polypeptides, there are many nucleotide sequences which are involved in control and regulation of gene expression.

The human genome project is directed toward determining the complete sequence of the genome of the human organism. Although such a sequence would not correspond to the sequence of any specific individual, it would provide significant information as to the general organization and specific sequences contained within segments from particular individuals. It would also provide mapping information which is very useful for further detailed studies. However, the need for highly rapid, accurate, and inexpensive sequencing technology is nowhere more apparent than in a

2

demanding sequencing project such as this. To complete the sequencing of a human genome would require the determination of approximately 3×10^9 , or 3 billion base pairs.

The procedures typically used today for sequencing include the Sanger dideoxy method, see, e.g., Sanger et al. (1977) *Proc. Natl. Acad. Sci. USA*, 74:5463-5467, or the Maxam and Gilbert method, see, e.g., Maxam et al., (1980) *Methods in Enzymology*, 65:499-559. The Sanger method utilizes enzymatic elongation procedures with chain terminating nucleotides. The Maxam and Gilbert method uses chemical reactions exhibiting specificity of reaction to generate nucleotide specific cleavages. Both methods require a practitioner to perform a large number of complex manual manipulations. These manipulations usually require isolating homogeneous DNA fragments, elaborate and tedious preparing of samples, preparing a separating gel, applying samples to the gel, electrophoresing the samples into this gel, working up the finished gel, and analyzing the results of the procedure.

Thus, a less expensive, highly reliable, and labor efficient means for sequencing biological macromolecules is needed. A substantial reduction in cost and increase in speed of nucleotide sequencing would be very much welcomed. In particular, an automated system would improve the reproducibility and accuracy of procedures. The present invention satisfies these and other needs.

SUMMARY OF THE INVENTION

The present invention provides improved methods useful for de novo sequencing of an unknown polymer sequence, for verification of known sequences, for fingerprinting polymers, and for mapping homologous segments within a sequence. By reducing the number of manual manipulations required and automating most of the steps, the speed, accuracy, and reliability of these procedures are greatly enhanced.

The production of a substrate having a matrix of positionally defined regions with attached reagents exhibiting known recognition specificity can be used for the sequence analysis of a polymer. Although most directly applicable to sequencing, the present invention is also applicable to fingerprinting, mapping, and general screening of specific interactions. The VLSIPS™ Technology (Very Large Scale Immobilized Polymer Synthesis) substrates will be applied to evaluating other polymers, e.g., carbohydrates, polypeptides, hydrocarbon synthetic polymers, and the like. For these non-polynucleotides, the sequence specific reagents will usually be antibodies specific for a particular subunit sequence.

According to one aspect of the masking technique, the invention provides an ordered method for forming a plurality of polymer sequences by sequential addition of reagents comprising the step of serially protecting and deprotecting portions of the plurality of polymer sequences for addition of other portions of the polymer sequences using a binary synthesis strategy.

The present invention also provides a means to automate sequencing manipulations. The automation of the substrate production method and of the scan and analysis steps minimizes the need for human intervention. This simplifies the tasks and promotes reproducibility.

The present invention provides a composition comprising a plurality of positionally distinguishable sequence specific reagents attached to a solid substrate, which reagents are capable of specifically binding to a predetermined subunit sequence of a preselected multi-subunit length having at

US 6,355,432 B1

3

least three subunits, said reagents representing substantially all possible sequences of said preselected length. In some embodiments, the subunit sequence is a polynucleotide or a polypeptide, in others the preselected multi-subunit length is five subunits and the subunit sequence is a polynucleotide sequence. In other embodiments, the specific reagent is an oligonucleotide of at least about five nucleotides. Alternatively, the specific reagent is a monoclonal antibody. Usually the specific reagents are all attached to a single solid substrate, and the reagents comprise about 3000 different sequences. In other embodiments, the reagents represents at least about 25% of the possible subsequences of said preselected length. Usually, the reagents are localized in regions of the substrate having a density of at least 25 regions per square centimeter, and often the substrate has a surface area of less than about 4 square centimeters.

The present invention also provides methods for analyzing a sequence of a polynucleotide or a polypeptide, said method comprising the step of:

- a) exposing said polynucleotide or polypeptide to a composition as described.

It also provides useful methods for identifying or comparing a target sequence with a reference, said method comprising the step of:

- a) exposing said target sequence to a composition as described;
- b) determining the pattern of positions of the reagents which specifically interact with the target sequence; and
- c) comparing the pattern with the pattern exhibited by the reference when exposed to the composition.

The present invention also provides methods for sequencing a segment of a polynucleotide comprising the steps of:

- a) combining:
 - i) a substrate comprising a plurality of chemically synthesized and positionally distinguishable oligonucleotides capable of recognizing defined oligonucleotide sequences; and
 - ii) a target polynucleotide; thereby forming high fidelity matched duplex structures of complementary subsequences of known sequence; and
- b) determining which of said reagents have specifically interacted with subsequences in said target polynucleotide.

In one embodiment, the segment is substantially the entire length of said polynucleotide.

The invention also provides methods for sequencing a polymer, said method comprising the steps of:

- a) preparing a plurality of reagents which each specifically bind to a subsequence of preselected length;
- b) positionally attaching each of said reagents to one or more solid phase substrates, thereby producing substrates of positionally definable sequence specific probes;
- c) combining said substrates with a target polymer whose sequence is to be determined; and
- d) determining which of said reagents have specifically interacted with subsequences in said target polymer.

In one embodiment, the substrates are beads. Preferably, the plurality of reagents comprise substantially all possible subsequences of said preselected length found in said target. In another embodiment, the solid phase substrate is a single substrate having attached thereto reagents recognizing substantially all possible subsequences of preselected length found in said target.

4

In another embodiment, the method further comprises the step of analyzing a plurality of said recognized subsequences to assemble a sequence of said target polymer. In a bead embodiment, at least some of the plurality of substrates have one subsequence specific reagent attached thereto, and the substrates are coded to indicate the sequence specificity of said reagent.

The present invention also embraces a method of using a fluorescent nucleotide to detect interactions with oligonucleotide probes of known sequence, said method comprising:

- a) attaching said nucleotide to a target unknown polynucleotide sequence, and
- b) exposing said target polynucleotide sequence to a collection of positionally defined oligonucleotide probes of known sequences to determine the sequences of said probes which interact with said target.

In a further refinement, an additional step is included of:

- a) collating said known sequences to determine the overlaps of said known sequences to determine the sequence of said target sequence.

A method of mapping a plurality of sequences relative to one another is also provided, the method comprising:

- a) preparing a substrate having a plurality of positionally attached sequence specific probes;
- b) exposing each of said sequences to said substrate, thereby determining the patterns of interaction between said sequence specific probes and said sequences; and
- c) determining the relative locations of said sequence specific probe interactions on said sequences to determine the overlaps and order of said sequences.

In one refinement, the sequence specific probes are oligonucleotides, applicable to where the target sequences are nucleic acid sequences.

In the nucleic acid sequencing application, the steps of the sequencing process comprise:

- a) producing a matrix substrate having known positionally defined regions of known sequence specific oligonucleotide probes;
- b) hybridizing a target polynucleotide to the positions on the matrix so that each of the positions which contain oligonucleotide probes complementary to a sequence on the target hybridize to the target molecule;
- c) detecting which positions have bound the target, thereby determining sequences which are found on the target; and
- d) analyzing the known sequences contained in the target to determine sequence overlaps and assembling the sequence of the target therefrom.

The enablement of the sequencing process by hybridization is based in large part upon the ability to synthesize a large number (e.g., to virtually saturate) of the possible overlapping sequence segments and distinguishing those probes which hybridize with fidelity from those which have mismatched bases, and to analyze a highly complex pattern of hybridization results to determine the overlap regions.

The detecting of the positions which bind the target sequence would typically be through a fluorescent label on the target. Although a fluorescent label is probably most convenient, other sorts of labels, e.g., radioactive, enzyme linked, optically detectable, or spectroscopic labels may be used. Because the oligonucleotide probes are positionally defined, the location of the hybridized duplex will directly translate to the sequences which hybridize. Thus, analysis of the positions provides a collection of subsequences found within the target sequence. These subsequences are matched

US 6,355,432 B1

5

with respect to their overlaps so as to assemble an intact target sequence.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 illustrates a flow chart for sequence, fingerprint, or mapping analysis.

FIGS. 2A-2M illustrate the proper function of a VLSIPS™ Technology nucleotide synthesis.

DESCRIPTION OF THE PREFERRED EMBODIMENTS

I. Overall Description

- A. general
- B. VLSIPS substrates
- C. binary masking
- D. applications
- E. detection methods and apparatus
- F. data analysis

II. Theoretical Analysis

- A. simple n-mer structure; theory
- B. complications
- C. non-polynucleotide embodiments

III. Polynucleotide Sequencing

- A. preparation of substrate matrix
- B. labeling target polynucleotide
- C. hybridization conditions
- D. detection; VLSIPS scanning
- E. analysis
- F. substrate reuse
- G. non-polynucleotide aspects

IV. Fingerprinting

- A. general
- B. preparation of substrate matrix
- C. labeling target nucleotides
- D. hybridization conditions
- E. detection; VLSIPS scanning
- F. analysis
- G. substrate reuse
- H. non-polynucleotide aspects

V. Mapping

- A. general
- B. preparation of substrate matrix
- C. labeling
- D. hybridization/specific interaction
- E. detection
- F. analysis
- G. substrate reuse
- H. non-polynucleotide aspects

VI. Additional Screening

- A. specific interactions
- B. sequence comparisons
- C. categorizations
- D. statistical correlations

VII. Formation of Substrate

- A. instrumentation
- B. binary masking
- C. synthetic methods
- D. surface immobilization

6

VIII. Hybridization/Specific Interaction

- A. general
- B. important parameters

IX. Detection Methods

- A. labeling techniques
- B. scanning system

X. Data Analysis

- A. general
- B. hardware
- C. software

XI. Substrate Reuse

- A. removal of label
- B. storage and preservation
- C. processes to avoid degradation of oligomers

XII. Integrated Sequencing Strategy

- A. initial mapping strategy
- B. selection of smaller clones
- C. actual sequencing procedures

XIII. Commercial Applications

- A. sequencing
- B. fingerprinting
- C. mapping

I. OVERALL DESCRIPTION

A. General

The present invention relies in part on the ability to synthesize or attach specific recognition reagents at known locations on a substrate, typically a single substrate. In particular, the present invention provides the ability to prepare a substrate having a very high density matrix pattern of positionally defined specific recognition reagents. The reagents are capable of interacting with their specific targets while attached to the substrate, e.g., solid phase interactions, and by appropriate labeling of these targets, the sites of the interactions between the target and the specific reagents may be derived. Because the reagents are positionally defined, the sites of the interactions will define the specificity of each interaction. As a result, a map of the patterns of interactions with specific reagents on the substrate is convertible into information on the specific interactions taking place, e.g., the recognized features. Where the specific reagents recognize a large number of possible features, this system allows the determination of the combination of specific interactions which exist on the target molecule. Where the number of features is sufficiently large, the identical same combination, or pattern, of features is sufficiently unlikely that a particular target molecule may often be uniquely defined by its features. In the extreme, the features may actually be the subunit sequence of the target molecule, and a given target sequence may be uniquely defined by its combination of features.

In particular, the methodology is applicable to sequencing polynucleotides. The specific sequence recognition reagents will typically be oligonucleotide probes which hybridize with specificity to subsequences found on the target sequence. A sufficiently large number of those probes allows the fingerprinting of a target polynucleotide or the relative mapping of a collection of target polynucleotides, as described in greater detail below.

In the high resolution fingerprinting provided by a saturating collection of probes which include all possible subsequences of a given size, e.g., 10-mers, collating of all the subsequences and determination of specific overlaps will be derived and the entire sequence can usually be reconstructed.

US 6,355,432 B1

7

Although a polynucleotide sequence analysis is a preferred embodiment, for which the specific reagents are most easily accessible, the invention is also applicable to analysis of other polymers, including polypeptides, carbohydrates, and synthetic polymers, including α -, β -, and ω -amino acids, polyurethanes, polyesters, polycarbonates, polyureas, polyamides, polyethyleneimines, polyarylene sulfides, polysiloxanes, polyimides, polyacetates, and mixed polymers. Various optical isomers, e.g., various D- and L-forms of the monomers, may be used.

Sequence analysis will take the form of complete sequence determination, to the level of the sequence of individual subunits along the entire length of the target sequence. Sequence analysis also takes the form of sequence homology, e.g., less than absolute subunit resolution, where "similarity" in the sequence will be detectable, or the form of selective sequences of homology interspersed at specific or irregular locations.

In either case, the sequence is determinable at selective resolution or at particular locations. Thus, the hybridization method will be useful as a means for identification, e.g., a "fingerprint", much like a Southern hybridization method is used. It is also useful to map particular target sequences.

B. VLSIPS™ Technology

The invention is enabled by the development of technology to prepare substrates on which specific reagents may be either positionally attached or synthesized. In particular, the very large scale immobilized polymer synthesis (VLSIPS™) technology allows for the very high density production of an enormous diversity of reagents mapped out in a known matrix pattern on a substrate. These reagents specifically recognize subsequences in a target polymer and bind thereto, producing a map of positionally defined regions of interaction. These map positions are convertible into actual features recognized, and thus would be present in the target molecule of interest.

As indicated, the sequence specific recognition reagents will often be oligonucleotides which hybridize with fidelity and discrimination to the target sequence. For use with other polymers, monoclonal or polyclonal antibodies having high sequence specificity will often be used.

In the generic sense, the VLSIPS technology allows the production of a substrate with a high density matrix of positionally mapped regions with specific recognition reagents attached at each distinct region. By use of protective groups which can be positionally removed, or added, the regions can be activated or deactivated for addition of particular reagents or compounds. Details of the protection are described below and in related Pirrung et al. (1992) U.S. Pat. No. 5,143,854. In a preferred embodiment, photosensitive protecting agents will be used and the regions of activation or deactivation may be controlled by electro-optical and optical methods, similar to many of the processes used in semiconductor wafer and chip fabrication.

In the nucleic acid nucleotide sequencing application, a VLSIPS substrate is synthesized having positionally defined oligonucleotide probes. See Pirrung et al. (1992) U.S. Pat. No. 5,143,854; and U.S. Ser. No. 07/624,120, now abandoned. By use of masking technology and photosensitive synthetic subunits, the VLSIPS apparatus allows for the stepwise synthesis of polymers according to a positionally defined matrix pattern. Each oligonucleotide probe will be synthesized at known and defined positional locations on the substrate. This forms a matrix pattern of known relationship between position and specificity of interaction. The VLSIPS technology allows the production of a very large number of different oligonucleotide probes to be simultaneously and

8

automatically synthesized including numbers in excess of about 10^2 , 10^3 , 10^4 , 10^5 , 10^6 , or even more, and at densities of at least about 10^2 , $10^3/\text{cm}^2$, $10^4/\text{cm}^2$, $10^5/\text{cm}^2$ and up to $10^6/\text{cm}^2$ or more. This application discloses methods for synthesizing polymers on a silicon or other suitably derivatized substrate, methods and chemistry for synthesizing specific types of biological polymers on those substrates, apparatus for scanning and detecting whether interaction has occurred at specific locations on the substrate, and various other technologies related to the use of a high density very large scale immobilized polymer substrate. In particular, sequencing, fingerprinting, and mapping applications are discussed herein in detail, though related technologies are described in simultaneously filed applications U.S. Ser. No. 07/624,120, now abandoned; and U.S. Ser. No. 07/517,659; Dower et al. (1995) U.S. Pat. No. 5,427,908, each of which is hereby incorporated herein by reference.

In other embodiments, antibody probes will be generated which specifically recognize particular subsequences found on a polymer. Antibodies would be generated which are specific for recognizing a three contiguous amino acid sequence, and monoclonal antibodies may be preferred. Optimally, these antibodies would not recognize any sequences other than the specific three amino acid stretch desired and the binding affinity should be insensitive to flanking or remote sequences found on a target molecule. Likewise, antibodies specific for particular carbohydrate linkages or sequences will be generated. A similar approach could be used for preparing specific reagents which recognize other polymer subunit sequences. These reagents would typically be site specifically localized to a substrate matrix pattern where the regions are closely packed.

These reagents could be individually attached at specific sites on the substrate in a matrix by an automated procedure where the regions are positionally targeted by some other specific mechanism, e.g., one which would allow the entire collection of reagents to be attached to the substrate in a single reaction. Each reagent could be separately attached to a specific oligonucleotide sequence by an automated procedure. This would produce a collection of reagents where, e.g., each monoclonal antibody would have a unique oligonucleotide sequence attached to it. By virtue of a VLSIPS substrate which has different complementary oligonucleotides synthesized on it, each monoclonal antibody would specifically be bound only at that site on the substrate where the complementary oligonucleotide has been synthesized. A crosslinking step would fix the reagent to the substrate. See, e.g., Dattagupta et al. (1985) U.S. Pat. No. 4,542,102 and (1987) U.S. Pat. No. 4,713,326; and Chatterjee, M. et al. (1990) *J. Am. Chem. Soc.* 112:6397-6399, which are hereby incorporated herein by reference. This allows a high density positionally specific collection of specific recognition reagents, e.g., monoclonal antibodies, to be immobilized to a solid substrate using an automated system.

The regions which define particular reagents will usually be generated by selective protecting groups which may be activated or deactivated. Typically the protecting group will be bound to a monomer subunit or spatial region, and can be spatially affected by an activator, such as electromagnetic radiation. Examples of protective groups with utility herein include nitroveratryl oxycarbonyl (NVOC), nitrobenzyl oxycarbonyl (NBOC), dimethyl dimethoxy benzyloxy carbonyl, 5-bromo-7-nitroindolyl, O-hydroxy- α -methyl cinnamoyl, and 2-oxymethylene anthraquinone. Examples of activators include ion beams, electric fields, magnetic fields, electron beams, x-ray, and other forms of electromagnetic radiation.

US 6,355,432 B1

9

C. Binary Masking

In fact, the means for producing a substrate useful for these techniques are explained in Pirrung et al. (1992) U.S. Pat. No. 5,143,854, which is hereby incorporated herein by reference. However, there are various particular ways to optimize the synthetic processes. Many of these methods are described in Ser. No. 07/624,120, now abandoned.

Briefly, the binary synthesis strategy refers to an ordered strategy for parallel synthesis of diverse polymer sequences by sequential addition of reagents which may be represented by a reactant matrix, and a switch matrix, the product of which is a product matrix. A reactant matrix is a 1xn matrix of the building blocks to be added. The switch matrix is all or a subset of the binary numbers from 1 to n arranged in columns. In preferred embodiments, a binary strategy is one in which at least two successive steps illuminate half of a region of interest on the substrate. In most preferred embodiments, binary synthesis refers to a synthesis strategy which also factors a previous addition step. For example, a strategy in which a switch matrix for a masking strategy halves regions that were previously illuminated, illuminating about half of the previously illuminated region and protecting the remaining half (while also protecting about half of previously protected regions and illuminating about half of previously protected regions). It will be recognized that binary rounds may be interspersed with non-binary rounds and that only a portion of a substrate may be subjected to a binary scheme, but will still be considered to be a binary masking scheme within the definition herein. A binary "masking" strategy is a binary synthesis which uses light to remove protective groups from materials for addition of other materials such as nucleotides or amino acids.

In particular, this procedure provides a simplified and highly efficient method for saturating all possible sequences of a defined length polymer. This masking strategy is also particularly useful in producing all possible oligonucleotide sequence probes of a given length.

D. Applications

The technology provided by the present invention has very broad applications. Although described specifically for polynucleotide sequences, similar sequencing, fingerprinting, mapping, and screening procedures can be applied to polypeptide, carbohydrate, or other polymers. In particular, the present invention may be used to completely sequence a given target sequence to subunit resolution. This may be for de novo sequencing, or may be used in conjunction with a second sequencing procedure to provide independent verification. See, e.g., (1988) *Science* 242:1245. For example, a large polynucleotide sequence defined by either the Maxam and Gilbert technique or by the Sanger technique may be verified by using the present invention.

In addition, by selection of appropriate probes, a polynucleotide sequence can be fingerprinted. Fingerprinting is a less detailed sequence analysis which usually involves the characterization of a sequence by a combination of defined features. Sequence fingerprinting is particularly useful because the repertoire of possible features which can be tested is virtually infinite. Moreover, the stringency of matching is also variable depending upon the application. A Southern Blot analysis may be characterized as a means of simple fingerprint analysis.

Fingerprinting analysis may be performed to the resolution of specific nucleotides, or may be used to determine homologies, most commonly for large segments. In particular, an array of oligonucleotide probes of virtually any workable size may be positionally localized on a matrix and used to probe a sequence for either absolute comple-

10

mentary matching, or homology to the desired level of stringency using selected hybridization conditions.

In addition, the present invention provides means for mapping analysis of a target sequence or sequences. Mapping will usually involve the sequential ordering of a plurality of various sequences, or may involve the localization of a particular sequence within a plurality of sequences. This may be achieved by immobilizing particular large segments onto the matrix and probing with a shorter sequence to determine which of the large sequences contain that smaller sequence. Alternatively, relatively shorter probes of known or random sequence may be immobilized to the matrix and a map of various different target sequences may be determined from overlaps. Principles of such an approach are described in some detail by Evans et al. (1989) "Physical Mapping of Complex Genomes by Cosmid Multiplex Analysis," *Proc. Natl. Acad. Sci. USA* 86:5030-5034; Michiels et al. (1987) "Molecular Approaches to Genome Analysis: A Strategy for the Construction of Ordered Overlap Clone Libraries," *CABIOS* 3:203-210; Olsen et al. (1986) "Random-Clone Strategy for Genomic Restriction Mapping in Yeast," *Proc. Natl. Acad. Sci. USA* 83:7826-7830; Craig, et al. (1990) "Ordering of Cosmid Clones Covering the Herpes Simplex Virus Type 1 (HSV-1) Genome: A Test Case for Fingerprinting by Hybridization," *Nuc. Acids Res.* 18:2653-2660; and Coulson, et al. (1986) "Toward a Physical Map of the Genome of the Nematode *Caenorhabditis elegans*," *Proc. Natl. Acad. Sci. USA* 83:7821-7825; each of which is hereby incorporated herein by reference.

Fingerprinting analysis also provides a means of identification. In addition to its value in apprehension of criminals from whom a biological sample, e.g., blood, has been collected, fingerprinting can ensure personal identification for other reasons. For example, it may be useful for identification of bodies in tragedies such as fire, flood, and vehicle crashes. In other cases the identification may be useful in identification of persons suffering from amnesia, or of missing persons. Other forensics applications include establishing the identity of a person, e.g., military identification "dog tags", or may be used in identifying the source of particular biological samples. Fingerprinting technology is described, e.g., in Carrano, et al. (1989) "A High-Resolution, Fluorescence-Based, Semi-automated method for DNA Fingerprinting," *Genomics* 4: 129-136, which is hereby incorporated herein by reference. See, e.g., table I, for nucleic acid applications, and corresponding applications may be accomplished using polypeptides.

TABLE I

VLSIPS™ TECHNOLOGY IN NUCLEIC ACIDS

- | | |
|-----|---|
| I. | Construction of Chips |
| II. | Applications |
| A. | Sequencing |
| 1. | Primary sequencing |
| 2. | Secondary sequencing (sequence checking) |
| 3. | Large scale mapping |
| 4. | Fingerprinting |
| B. | Duplex/Triplex formation |
| 1. | Antisense |
| 2. | Sequence specific function modulation
(e.g. promoter inhibition) |
| C. | Diagnosis |
| 1. | Genetic markers |
| 2. | Type markers |
| a. | Blood donors |
| b. | Tissue transplants |

US 6,355,432 B1

11

TABLE I-continued

VLSIPS™ TECHNOLOGY IN NUCLEIC ACIDS	
D. Microbiology	
1. Clinical microbiology	
2. Food microbiology	
III. Instrumentation	
A. Chip machines	
B. Detection	
IV. Software Development	
A. Instrumentation software	
B. Data reduction software	
C. Sequence analysis software	

The fingerprinting analysis may be used to perform various types of genetic screening. For example, a single substrate may be generated with a plurality of screening probes, allowing for the simultaneous genetic screening for a large number of genetic markers. Thus, prenatal or diagnostic screening can be simplified, economized, and made more generally accessible.

In addition to the sequencing, fingerprinting, and mapping applications, the present invention also provides means for determining specificity of interaction with particular sequences. Many of these applications were described in Ser. No. 07/362,901, now abandoned, Pirrung et al. (1992) U.S. Pat. No. 5,143,854; Ser. No. 07/435,316, and Ser. No. 07/612,671.

E. Detection Methods and Apparatus

An appropriate detection method applicable to the selected labeling method can be selected. Suitable labels include radionucleotides, enzymes, substrates, cofactors, inhibitors, magnetic particles, heavy metal atoms, and particularly fluorescers, chemiluminescers, and spectroscopic labels. Patents teaching the use of such labels include U.S. Pat. Nos. 3,817,837; 3,850,752; 3,939,350; 3,996,345; 4,277,437; 4,275,149; and 4,366,241.

With an appropriate label selected, the detection system best adapted for high resolution and high sensitivity detection may be selected. As indicated above, an optically detectable system, e.g., fluorescence or chemiluminescence would be preferred. Other detection systems may be adapted to the purpose, e.g., electron microscopy, scanning electron microscopy (SEM), scanning tunneling electron microscopy (STEM), infrared microscopy, atomic force microscopy (AFM), electrical conductance, and image plate transfer.

With a detection method selected, an apparatus for scanning the substrate will be designed. Apparatus, as described in Ser. No. 07/362,901, now abandoned; or Pirrung et al. (1992) U.S. Pat. No. 5,143,854; or Ser. No. 07/624,120, now abandoned, are particularly appropriate. Design modifications may also be incorporated therein.

F. Data Analysis

Data is analyzed by processes similar to those described below in the section describing theoretical analysis. More efficient algorithms will be mathematically devised, and will usually be designed to be performed on a computer. Various computer programs which may more quickly or efficiently make measurement samples and distinguish signal from noise will also be devised. See, particularly, Ser. No. 07/624,120, now abandoned.

The initial data resulting from the detection system is an array of data indicative of fluorescent intensity versus location on the substrate. The data are typically taken over regions substantially smaller than the area in which synthesis of a given polymer has taken place. Merely by way of example, if polymers were synthesized in squares on the substrate having dimensions of 500 microns by 500 microns, the data may be taken over regions having dimensions of 5 microns by 5 microns. In most preferred embodiments, the

12

regions over which fluorescence data are taken across the substrate are less than about 1/2 the area of the regions in which individual polymers are synthesized, preferably less than 1/10 the area in which a single polymer is synthesized, and most preferably less than 1/100 the area in which a single polymer is synthesized. Hence, within any area in which a given polymer has been synthesized, a large number of fluorescence data points are collected.

A plot of number of pixels versus intensity for a scan should bear a rough resemblance to a bell curve, but spurious data are observed, particularly at higher intensities. Since it is desirable to use an average of fluorescent intensity over a given synthesis region in determining relative binding affinity, these spurious data will tend to undesirably skew the data.

Accordingly, in one embodiment of the invention the data are corrected for removal of these spurious data points, and an average of the data points is thereafter utilized in determining relative binding efficiency. In general the data are fitted to a base curve and statistical measures are used to remove spurious data.

In an additional analytical tool, various degeneracy reducing analogues may be incorporated in the hybridization probes. Various aspects of this strategy are described, e.g., in Macevitz, S. (1990) PCT publication number WO 90/04652, which is hereby incorporated herein by reference.

II. THEORETICAL ANALYSIS

The principle of the hybridization sequencing procedure is based, in part, upon the ability to determine overlaps of short segments. The VLSIPS technology provides the ability to generate reagents which will saturate the possible short subsequence recognition possibilities. The principle is most easily illustrated by using a binary sequence, such as a sequence of zeros and ones. Once having illustrated the application to a binary alphabet, the principle may easily be understood to encompass three letter, four letter, five or more letter, even 20 letter alphabets. A theoretical treatment of analysis of subsequence information to reconstruction of a target sequence is provided, e.g., in Lysov, Yu., et al. (1988) *Doklady Akademi. Nauk. SSR* 303:1508-1511; Khrapko K., et al. (1989) *FEBS Letters* 256:118-122; Pevzner, P. (1989) *J. of Biomolecular Structure and Dynamics* 7:63-69; and Drmanac, R. et al. (1989) *Genomics* 4:114-128; each of which is hereby incorporated herein by reference.

The reagents for recognizing the subsequences will usually be specific for recognizing a particular polymer subsequence anywhere within a target polymer. It is preferable that conditions may be devised which allow absolute discrimination between high fidelity matching and very low levels of mismatching. The reagent interaction will preferably exhibit no sensitivity to flanking sequences, to the subsequence position within the target, or to any other remote structure within the sequence. For polynucleotide sequencing, the specific reagents can be oligonucleotide probes; for polypeptides and carbohydrates, antibodies will be useful reagents. Antibody reagents should also be useful for other types of polymers.

A. Simple n-mer Structure: Theory

1. Simple Two Letter Alphabet: Example

A simple example is presented below of how a sequence of ten digits comprising zeros and ones would be sequenceable using short segments of five digits. For example, consider the sample ten digit sequence:

1010011100.

A VLSIPS™ Technology substrate could be constructed, as discussed elsewhere, which would have reagents attached in

US 6,355,432 B1

13

a defined matrix pattern which specifically recognize each of the possible five digit sequences of ones and zeros. The number of possible five digit subsequences is $2^5=32$. The number of possible different sequences 10 digits long is $2^{10}=1,024$. The five contiguous digit subsequences within a ten digit sequence number six, i.e., positioned at digits 1-5, 2-6, 3-7, 4-8, 5-9, and 6-10. It will be noted that the specific order of the digits in the sequence is important and that the order is directional, e.g., running left to right versus right to left. The first five digit sequence contained in the target sequence is 10100. The second is 01001, the third is 10011, the fourth is 00111, the fifth is 01110, and the sixth is 11100.

The VLSIPS™ substrate would have a matrix pattern of positionally attached reagents which recognize each of the different 5-mer subsequences. Those reagents which recognize each of the 6 contained 5-mers will bind the target, and a label allows the positional determination of where the sequence specific interaction has occurred. By correlation of the position in the matrix pattern, the corresponding bound subsequences can be determined.

In the above-mentioned sequence, six different 5-mer sequences would be determined to be present. They would be:

```
10100
01001
10011
00111
01110
11100
```

Any sequence which contains the first five digit sequence, 10100, already narrows the number of possible sequences (e.g., from 1024 possible sequences) which contain it to less than about 192 possible sequences.

This 192 is derived from the observation that with the subsequence 10100 at the far left of the sequence, in positions 1-5, there are only 32 possible sequences. Likewise, for that particular subsequence in positions 2-6, 3-7, 4-8, 5-9, and 6-10. So, to sum up all of the sequences that could contain 10100, there are 32 for each position and 6 positions for a total of about 192 possible sequences. However, some of these 10 digit sequences will have been counted twice. Thus, by virtue of containing the 10100 subsequence, the number of possible 10-mer sequences has been decreased from 1024 sequences to less than about 192 sequences.

In this example, not only do we know that the sequence contains 10100, but we also know that it contains the second five character sequence, 01001. By virtue of knowing that the sequence contains 10100, we can look specifically to determine whether the sequence contains a subsequence of five characters which contains the four leftmost digits plus a next digit to the left. For example, we would look for a sequence of X1010, but we find that there is none. Thus, we know that the 10100 must be at the left end of the 10-mer. We would also look to see whether the sequence contains the rightmost four digits plus a next digit to the right, e.g., 0100X. We find that the sequence also contains the sequence 01001, and that X is a 1. Thus, we know at least that our target sequence has an overlap of 0100 and has the left terminal sequence 101001.

Applying the same procedure to the second 5-mer, we also know that the sequence must include a sequence of five

14

digits having the sequence 1001Y where Y must be either 0 or 1. We look through the fragments and we see that we have a 10011 sequence within our target, thus Y is also 1. Thus, we would know that our sequence has a sequence of the first seven being 1010011.

Moving to the next 5-mer, we know that there must be a sequence of 0011Z, where Z must be either 0 or 1. We look at the fragments produced above and see that the target sequence contains a 00111 subsequence and Z is 1. Thus, we know the sequence must start with 10100111.

The next 5-mer must be of the sequence 0111W where W must be 0 or 1. Again, looking up at the fragments produced, we see that the target sequence contains a 01110 subsequence, and W is a 0. Thus, our sequence to this point is 101001110. We know that the last 5-mer must be either 11100 or 11101. Looking above, we see that it is 11100 and that must be the last of our sequence. Thus, we have determined that our sequence must have been 1010011100.

However, it will be recognized from the example above with the sequences provided therein, that the sequence analysis can start with any known positive probe subsequence. The determination may be performed by moving linearly along the sequence checking the known sequence with a limited number of next positions. Given this possibility, the sequence may be determined, besides by scanning all possible oligonucleotide probe positions, by specifically looking only where the next possible positions would be. This may increase the complexity of the scanning but may provide a longer time span dedicated towards scanning and detecting specific positions of interest relative to other sequence possibilities. Thus, the scanning apparatus could be set up to work its way along a sequence from a given contained oligonucleotide to only look at those positions on the substrate which are expected to have a positive signal.

It is seen that given a sequence, it can be de-constructed into n-mers to produce a set of internal contiguous subsequences. From any given target sequence, we would be able to determine what fragments would result. The hybridization sequence method depends, in part, upon being able to work in the reverse, from a set of fragments of known sequences to the full sequence. In simple cases, one is able to start at a single position and work in either or both directions towards the ends of the sequence as illustrated in the example.

The number of possible sequences of a given length increases very quickly with the length of that sequence. Thus, a 10-mer of zeros and ones has 1024 possibilities, a 12-mer has 4096. A 20-mer has over a million possibilities, and a 30-mer has over a billion. However, a given 30-mer has, at most, 26 different internal 5-mer sequences. Thus, a 30 character target sequence having over a million possible sequences can be substantially defined by only 26 different 5-mers. It will be recognized that the probe oligonucleotides will preferably, but need not necessarily, be of identical length, and that the probe sequences need not necessarily be contiguous in that the overlapping subsequences need not differ by only a single subunit. Moreover, each position of the matrix pattern need not be homogeneous, but may actually contain a plurality of probes of known sequence. In addition, although all of the possible subsequence specifications would be preferred, a less than full set of sequence specifications could be used. In particular, although a substantial fraction will preferably be at least about 70%, it may be less than that. About 20% would be preferred, more preferably at least about 30% would be desired. Higher percentages would be especially preferred.

IAFP00000674

US 6,355,432 B1

15

2. Example of Four Letter Alphabet

A four letter alphabet may be conceptualized in at least two different ways from the two letter alphabet. One way is to consider the four possible values at each position and to analogize in a similar fashion to the binary example each of the overlaps. A second way is to group the binary digits into groups.

Using the first means, the overlap comparisons are performed with a four letter alphabet rather than a two letter alphabet. Then, in contrast to the binary system with 10 positions where $2^{10}=1024$ possible sequences, in a 4-character alphabet with 10 positions, there will actually be $4^{10}=1,048,576$ possible sequences. Thus, the complexity of a four character sequence has a much larger number of possible sequences compared to a two character sequence. Note, however, that there are still only 6 different internal 5-mers. For simplicity, we shall examine a 5 character string with 3 character subsequences. Instead of only 1 and 0, the characters may be designated, e.g., A, C, G, and T. Let us take the sequence GGCTA. The 3-mer subsequences are:

GGC
GCT
CTA

Given these subsequences, there is one sequence, or at most only a few sequences which would produce that combination of subsequences, i.e., GGCTA.

Alternatively, with a four character universe, the binary system can be looked at in pairs of digits. The pairs would be 00, 01, 10, and 11. In this manner, the earlier used sequence 1010011100 is looked at as 10,10,01,11,00. Then the first character of two digits is selected from the possible universe of the four representations 00, 01, 10, and 11. Then a probe would be in an even number of digits, e.g., not five digits, but, three pairs of digits or six digits. A similar comparison is performed and the possible overlaps determined. The 3-pair subsequences are:

10, 10, 01
10, 01, 11
01, 11, 00

and the overlap reconstruction produces 10,10,01,11,00.

The latter of the two conceptual views of the 4 letter alphabet provides a representation which is similar to what would be provided in a digital computer. The applicability to a four nucleotide alphabet is easily seen by assigning, e.g., 00 to A, 01 to C, 10 to G, and 11 to T. And, in fact, if such a correspondence is used, both examples for the 4 character sequences can be seen to represent the same target sequence. The applicability of the hybridization method and its analysis for determining the ultimate sequence is easily seen if A is the representation of adenine, C is the representation of cytosine, G is the representation of guanine, and T is the representation of thymine or uracil.

3. Generalization to m-letter Alphabet

This reconstruction process may be applied to polymers of virtually any number of possible characters in the alphabet, and for virtually any length sequence to be sequenced, though limitations, as discussed below, will limit its efficiency at various extremes of length. It will be recognized that the theory can be applied to a large diversity of systems where sequence is important.

16

For example, the method could be applied to sequencing of a polypeptide. A polypeptide can have any of twenty natural amino acid possibilities at each position. A twenty letter alphabet is amenable to sequencing by this method so long as reagents exist for recognizing shorter subsequences therein. A preferred reagent for achieving that goal would be a set of monoclonal antibodies each of which recognizes a specific three contiguous amino acid subsequence. A complete set of antibodies which recognize all possible subsequences of a given length, e.g., 3 amino acids, and preferably with a uniform affinity, would be $20^3=8000$ reagents.

It will also be recognized that each target sequence which is recognized by the specific reagents need not have homogeneous termini. Thus, fragments of the entire target sequence will also be useful for hybridizing appropriate subsequences. It is, however, preferable that there not be a significant amount of labeled homogeneous contaminating extraneous sequences. This constraint does usually require the purification of the target molecule to be sequenced, but a specific label technique would dispense with a purification requirement if the unlabeled extraneous sequences do not interfere with the labeled sequences.

In addition, conformational effects of target polypeptide folding may, in certain embodiments, be negligible if the polypeptide is fragmented into sufficiently small peptides, or if the interaction is performed under conditions where conformation, but not specific interaction, is disrupted.

B. Complications

Two obvious complications exist with the method of sequence analysis by hybridization. The first results from a probe of inappropriate length while the second relates to internally repeated sequences.

The first obvious complication is a problem which arises from an inappropriate length of recognition sequence, which causes problems with the specificity of recognition. For example, if the recognized sequence is too short, every sequence which is utilized will be recognized by every probe sequence. This occurs, e.g., in a binary system where the probes are each of sequences which occur relatively frequently, e.g., a two character probe for the binary system. Each possible two character probe would be expected to appear $\frac{1}{4}$ of the time in every single two character position. Thus, the above sequence example would be recognized by each of the 00, 10, 01, and 11. Thus, the sequence information is virtually lost because the resolution is too low and each recognition reagent specifically binds at multiple sites on the target sequence.

The number of different probes which bind to a target depends on the relationship between the probe length and the target length. At the extreme of short probe length, the just mentioned problem exists of excessive redundancy and lack of resolution. The lack of stability in recognition will also be a problem with extremely short probes. At the extreme of long probe length, each entire probe sequence is on a different position of a substrate. However, a problem arises from the number of possible sequences, which goes up dramatically with the length of the sequence. Also, the specificity of recognition begins to decrease as the contribution to binding by any particular subunit may become sufficiently low that the system fails to distinguish the fidelity of recognition. Mismatched hybridization may be a problem with the polynucleotide sequencing applications, though the fingerprinting and mapping applications may not be so strict in their fidelity requirements. As indicated above, a thirty position binary sequence has over a million possible sequences, a number which starts to become unreasonably large in its required number of different sequences, even

US 6,355,432 B1

17

though the target length is still very short. Preparing a substrate with all sequence possibilities for a long target may be extremely difficult due to the many different oligomers which must be synthesized.

The above example illustrates how a long target sequence may be reconstructed with a reasonably small number of shorter subsequences. Since the present day resolution of the regions of the substrate having defined oligomer probes attached to the substrate approaches about 10 microns by 10 microns for resolvable regions, about 10^6 , or 1 million, positions can be placed on a one centimeter square substrate. However, high resolution systems may have particular disadvantages which may be outweighed using the lower density substrate matrix pattern. For this reason, a sufficiently large number of probe sequences can be utilized so that any given target sequence may be determined by hybridization to a relatively small number of probes.

A second complication relates to convergence of sequences to a single subsequence. This will occur when a particular subsequence is repeated in the target sequence. This problem can be addressed in at least two different ways. The first, and simpler way, is to separate the repeat sequences onto two different targets. Thus, each single target will not have the repeated sequence and can be analyzed to its end. This solution, however, complicates the analysis by requiring that some means for cutting at a site between the repeats can be located. Typically a careful sequencer would want to have two intermediate cut points so that the intermediate region can also be sequenced in both directions across each of the cut points. This problem is inherent in the hybridization method for sequencing but can be minimized by using a longer known probe sequence so that the frequency of probe repeats is decreased.

Knowing the sequence of flanking sequences of the repeat will simplify the use of polymerase chain reaction (PCR) or a similar technique to further definitively determine the sequence between sequence repeats. Probes can be made to hybridize to those known sequences adjacent the repeat sequences, thereby producing new target sequences for analysis. See, e.g., Innis et al. (eds.) (1990) *PCR Protocols: A Guide to Methods and Applications*, Academic Press; and methods for synthesis of oligonucleotide probes, see, e.g., Gait (1984) *Oligonucleotide Synthesis: A Practical Approach*, IRL Press, Oxford.

Other means for dealing with convergence problems include using particular longer probes, and using degeneracy reducing analogues, see, e.g., Macevicz, S. (1990) PCT publication number WO 90/04652, which is hereby incorporated herein by reference. By use of stretches of the degeneracy reducing analogues with other probes in particular combinations, the number of probes necessary to fully saturate the possible oligomer probes is decreased. For example, with a stretch of 12-mers having the central 4-mer of degenerate nucleotides, in combination with all of the possible 8-mers, the collection numbers twice the number of possible 8-mers, e.g. $65,536 + 65,536 = 131,072$, but the population provides screening equivalent to all possible 12-mers.

By way of further explanation, all possible oligonucleotide 8-mers may be depicted in the fashion:

N1-N2-N3-N4-N5-N6-N7-N8,

in which there are $4^8 = 65,536$ possible 8-mers. As described in Ser. No. 07/624,120, now abandoned, producing all possible 8-mers requires $4 \times 8 = 32$ chemical binary synthesis steps to produce the entire matrix pattern of 65,536 8-mer possibilities. By incorporating degeneracy reducing

18

nucleotides, D's, which hybridize nonselectively to any corresponding complementary nucleotide, new oligonucleotides 12-mers can be made in the fashion:

N1-N2-N3-N4-D-D-D-D-N5-N6-N7-N8,

in which there are again, as above, only $4^8 = 65,536$ possible "12-mers", which in reality only have 8 different nucleotides.

However, it can be seen that each possible 12-mer probe could be represented by a group of the two 8-mer types. Moreover, repeats of less than 12 nucleotides would not converge, or cause repeat problems in the analysis. Thus, instead of requiring a collection of probes corresponding to all 12-mers, or $4^{12} = 16,777,216$ different 12-mers, the same information can be derived by making 2 sets of "8-mers" consisting of the typical 8-mer collection of $4^8 = 65,536$ and the "12-mer" set with the degeneracy reducing analogues, also requiring making $4^8 = 65,536$. The combination of the two sets, requires making $65,536 + 65,536 = 131,072$ different molecules, but giving the information of 16,777,216 molecules. Thus, incorporating the degeneracy reducing analogue decreases the number of molecules necessary to get 12-mer resolution by a factor of about 128-fold.

C. Non-polynucleotide Embodiments

The above example is directed towards a polynucleotide embodiment. This application is relatively easily achieved because the specific reagents will typically be complementary oligonucleotides, although in certain embodiments other specific reagents may be desired. For example, there may be circumstances where other than complementary base pairing will be utilized. The polynucleotide targets, will usually be single strand, but may be double or triple stranded in various applications. However, a triple stranded specific interaction might be sometimes desired, or a protein or other specific binding molecule may be utilized. For example, various promoter or DNA sequence specific binding proteins might be used, including, e.g., restriction enzyme binding domains, other binding domains, and antibodies. Thus, specific recognition reagents besides oligonucleotides may be utilized.

For other polymer targets, the specific reagents will often be polypeptides. These polypeptides may be protein binding domains from enzymes or other proteins which display specificity for binding. Usually an antibody molecule may be used, and monoclonal antibodies may be particularly desired. Classical methods may be applied for preparing antibodies, see, e.g., Harlow and Lane (1988) *Antibodies: A Laboratory Manual* Cold Spring Harbor Press, New York; and Goding (1986) *Monoclonal Antibodies: Principles and Practice* (2d Ed.) Academic Press, San Diego. Other suitable techniques for in vitro exposure of lymphocytes to the antigens or selection of libraries of antibody binding sites are described, e.g., in Huse et al. (1989) *Science* 246:1275-1281; and Ward et al. 91989) *Nature* 341:544-546, each of which is hereby incorporated herein by reference. Unusual antibody production methods are also described, e.g., in Hendricks et al. (1989) *BioTechnology* 7:1271-1274; and Hiatt et al. (1989) *Nature* 342:76-78, each of which is hereby incorporated herein by reference. Other molecules which may exhibit specific binding interaction may be useful for attachment to a VLSIPS substrate by various methods, including the caged biotin methods, see, e.g., Ser. No. 07/435,316, now abandoned, and Barrett et al. (1993) U.S. Pat. No. 5,252,743.

The antibody specific reagents should be particularly useful for the polypeptide, carbohydrate, and synthetic poly-

US 6,355,432 B1

19

mer applications. Individual specific reagents might be generated by an automated process to generate the number of reagents necessary to advantageously use the high density positional matrix pattern. In an alternative approach, a plurality of hybridoma cells may be screened for their ability to bind to a VLSIPS matrix possessing the desired sequences whose binding specificity is desired. Each cell might be individually grown up and its binding specificity determined by VLSIPS apparatus and technology. An alternative strategy would be to expose the same VLSIPS matrix to a polyclonal serum of high titer. By a successively large volume of serum and different animals, each region of the VLSIPS substrate would have attached to it a substantial number of antibody molecules with specificity of binding. The substrate, with non-covalently bound antibodies could be derivatized and the antibodies transferred to an adjacent second substrate in the matrix pattern in which the antibody molecules had attached to the first matrix. If the sensitivity of detection of binding interaction is sufficiently high, such a low efficiency transfer of antibody molecules may produce a sufficiently high signal to be useful for many purposes, including the sequencing applications.

In another embodiment, capillary forces may be used to transfer the selected reagents to a new matrix, to which the reagents would be positionally attached in the pattern of the recognized sequences. Or, the reagents could be transversely electrophoresed, magnetically transferred, or otherwise transported to a new substrate in their retained positional pattern.

III. POLYNUCLEOTIDE SEQUENCING

In principle, the making of a substrate having a positionally defined matrix pattern of all possible oligonucleotides of a given length involves a conceptually simple method of synthesizing each and every different possible oligonucleotide, and affixing them to a definable position. Oligonucleotide synthesis is presently mechanized and enabled by current technology, see, e.g., Ser. No. 07/362,901, now abandoned; Pirrung et al. (1992) U.S. Pat. No. 5,143,854; and instruments supplied by Applied Biosystems, Foster City, Calif.

A. Preparation of Substrate Matrix

The production of the collection of specific oligonucleotides used in polynucleotide sequencing may be produced in at least two different ways. Present technology certainly allows production of ten nucleotide oligomers on a solid phase or other synthesizing system. See, e.g., instrumentation provided by Applied Biosystems, Foster City, Calif. Although a single oligonucleotide can be relatively easily made, a large collection of them would typically require a fairly large amount of time and investment. For example, there are 4^{10} —1,048,576 possible ten nucleotide oligomers. Present technology allows making each and every one of them in a separate purified form though such might be costly and laborious.

Once the desired repertoire of possible oligomer sequences of a given length have been synthesized, this collection of reagents may be individually positionally attached to a substrate, thereby allowing a batchwise hybridization step. Present technology also would allow the possibility of attaching each and every one of these 10-mers to a separate specific position on a solid matrix. This attachment could be automated in any of a number of ways, particularly through the use of a caged biotin type linking. This would produce a matrix having each of different possible 10-mers.

A batchwise hybridization is much preferred because of its reproducibility and simplicity. An automated process of

20

attaching various reagents to positionally defined sites on a substrate is provided in Pirrung et al. (1992) U.S. Pat. No. 5,143,854; Ser. No. 07/624,120, now abandoned; and Barrett et al. (1993) U.S. Pat. No. 5,252,743; each of which is hereby incorporated herein by reference.

Instead of separate synthesis of each oligonucleotide, these oligonucleotides are conveniently synthesized in parallel by sequential synthetic processes on a defined matrix pattern as provided in Pirrung et al. (1992) U.S. Pat. No. 5,143,854; and Ser. No. 07/624,120, now abandoned, which are incorporated herein by reference. Here, the oligonucleotides are synthesized stepwise on a substrate at positionally separate and defined positions. Use of photosensitive blocking reagents allows for defined sequences of synthetic steps over the surface of a matrix pattern. By use of the binary masking strategy, the surface of the substrate can be positioned to generate a desired pattern of regions, each having a defined sequence oligonucleotide synthesized and immobilized thereto.

Although the prior art technology can be used to generate the desired repertoire of oligonucleotide probes, an efficient and cost effective means would be to use the VLSIPS technology described in Pirrung et al. (1992) U.S. Pat. No. 5,143,854 and Ser. No. 07/624,120, now abandoned. In this embodiment, the photosensitive reagents involved in the production of such a matrix are described below.

The regions for synthesis may be very small, usually less than about $100\text{ }\mu\text{m}\times 100\text{ }\mu\text{m}$, more usually less than about $50\text{ }\mu\text{m}\times 50\text{ }\mu\text{m}$. The photolithography technology allows synthetic regions of less than about $10\text{ }\mu\text{m}\times 10\text{ }\mu\text{m}$, about $3\text{ }\mu\text{m}\times 3\text{ }\mu\text{m}$, or less. The detection also may detect such sized regions, though larger areas are more easily and reliably measured.

At a size of about 30 microns by 30 microns, one million regions would take about 11 centimeters square or a single wafer of about 4 centimeters by 4 centimeters. Thus the present technology provides for making a single matrix of that size having all one million plus possible oligonucleotides. Region size is sufficiently small to correspond to densities of at least about 5 regions/cm², 20 regions/cm², 50 regions/cm², 100 regions/cm², and greater, including 300 regions/cm², 1000 regions/cm², 3K regions/cm², 10K regions/cm², 30K regions/cm², 100K regions/cm², 300K regions/cm² or more, even in excess of one million regions/cm².

Although the pattern of the regions which contain specific sequences is theoretically not important, for practical reasons certain patterns will be preferred in synthesizing the oligonucleotides. The application of binary masking algorithms for generating the pattern of known oligonucleotide probes is described in related Ser. No. 07/624,120, now abandoned, which was filed simultaneously with this application. By use of these binary masks, a highly efficient means is provided for producing the substrate with the desired matrix pattern of different sequences. Although the binary masking strategy allows for the synthesis of all lengths of polymers, the strategy may be easily modified to provide only polymers of a given length. This is achieved by omitting steps where a subunit is not attached.

The strategy for generating a specific pattern may take any of a number of different approaches. These approaches are well described in related application Ser. No. 07/624,120, now abandoned, and include a number of binary masking approaches which will not be exhaustively discussed herein. However, the binary masking and binary synthesis approaches provide a maximum of diversity with a minimum number of actual synthetic steps.

US 6,355,432 B1

21

The length of oligonucleotides used in sequencing applications will be selected on criteria determined to some extent by the practical limits discussed above. For example, if probes are made as oligonucleotides, there will be 65,536 possible eight nucleotide sequences. If a nine subunit oligonucleotide is selected, there are 262,144 possible permutations of sequences. If a ten-mer oligonucleotide is selected, there are 1,048,576 possible permutations of sequences. As the number gets larger, the required number of positionally defined subunits necessary to saturate the possibilities also increases. With respect to hybridization conditions, the length of the matching necessary to confer stability of the conditions selected can be compensated for. See, e.g., Kanehisa, M. (1984) *Nuc. Acids Res.* 12:203-213, which is hereby incorporated herein by reference.

Although not described in detail here, but below for oligonucleotide probes, the VLSIPS technology would typically use a photosensitive protective group on an oligonucleotide. Sample oligonucleotides are shown in FIG. 1. In particular, the photoprotective group on the nucleotide molecules may be selected from a wide variety of positive light reactive groups preferably including nitro aromatic compounds such as o-nitro-benzyl derivatives or benzylsulfonyl. See, e.g., Gait (1984) *Oligonucleotide Synthesis: A Practical Approach*, IRL Press, Oxford, which is hereby incorporated herein by reference. In a preferred embodiment, 6-nitro-veratryl oxycarbonyl (NVOC), 2-nitrobenzyl oxycarbonyl (NBOC), or α,α -dimethyl-dimethoxybenzyl oxycarbonyl (DEZ) is used. Photoremovable protective groups are described in, e.g., Patchornik (1970) *J. Amer. Chem. Soc.* 92:6333-6335; and Amit et al. (1974) *J. Organic Chem.* 39:192-196; each of which is hereby incorporated herein by reference.

A preferred linker for attaching the oligonucleotide to a silicon matrix is illustrated in FIG. 2. A more detailed description is provided below. A photosensitive blocked nucleotide may be attached to specific locations of unblocked prior cycles of attachments on the substrate and can be successively built up to the correct length oligonucleotide probe.

It should be noted that multiple substrates may be simultaneously exposed to a single target sequence where each substrate is a duplicate of one another or where, in combination, multiple substrates together provide the complete or desired subset of possible subsequences. This provides the opportunity to overcome a limitation of the density of positions on a single substrate by using multiple substrates. In the extreme case, each probe might be attached to a single bead or substrate and the beads sorted by whether there is a binding interaction. Those beads which do bind might be encoded to indicate the subsequence specificity of reagents attached thereto.

Then, the target may be bound to the whole collection of beads and those beads that have appropriate specific reagents on them will bind to the target. Then a sorting system may be utilized to sort those beads that actually bind the target from those that do not. This may be accomplished by presently available cell sorting devices or a similar apparatus. After the relatively small number of beads which have bound the target have been collected, the encoding scheme may be read off to determine the specificity of the reagent on the bead. An encoding system may include a magnetic system, a shape encoding system, a color encoding system, or a combination of any of these, or any other encoding system. Once again, with the collection of specific interactions that have occurred, the binding may be analyzed for sequence information, fingerprint information, or mapping information.

22

The parameters of polynucleotide sizes of both the probes and target sequences are determined by the applications and other circumstances. The length of the oligonucleotide probes used will depend in part upon the limitations of the VLSIPS technology to provide the number of desired probes. For example, in an absolute sequencing application, it is often useful to have virtually all of the possible oligonucleotides of a given length. As indicated above, there are 65,536 8-mers, 262,144 9-mers, 1,048,576 10-mers, 4,194,304 11-mers, etc. As the length of the oligomer increases the number of different probes which must be synthesized also increases at a rate of a factor of 4 for every additional nucleotide. Eventually the size of the matrix and the limitations in the resolution of regions in the matrix will reach the point where an increase in number of probes becomes disadvantageous. However, this sequencing procedure requires that the system be able to distinguish, by appropriate selection of hybridization and washing conditions, between binding of absolute fidelity and binding of complementary sequences containing mismatches. On the other hand, if the fidelity is unnecessary, this discrimination is also unnecessary and a significantly longer probe may be used. Significantly longer probes would typically be useful in fingerprinting or mapping applications.

The length of the probe is selected for a length that will allow the probe to bind with specificity to possible targets. The hybridization conditions are also very important in that they will determine how closely the homology of complementary binding will be detected. In fact, a single target may be evaluated at a number of different conditions to determine its spectrum of specificity for binding particular probes. This may find use in a number of other applications besides the polynucleotide sequencing fingerprinting or mapping. For example, it will be desired to determine the spectrum of binding affinities and specificities of cell surface antigens with binding by particular antibodies immobilized on the substrate surface, particularly under different interaction conditions. In a related fashion, different regions with reagents having differing affinities or levels of specificity may allow such a spectrum to be defined using a single incubation, where various regions, at a given hybridization condition, show the binding affinity. For example, fingerprint probes of various lengths, or with specific defined non-matches may be used. Unnatural nucleotides or nucleotides exhibiting modified specificity of complementary binding are described in greater detail in Maccvitz (1990) PCT pub. No. WO 90/04652; and see the section on modified nucleotides in the Sigma Chemical Company catalogue. B. Labeling Target Nucleotide

The label used to detect the target sequences will be determined, in part, by the detection methods being applied. Thus, the labeling method and label used are selected in combination with the actual detecting systems being used.

Once a particular label has been selected, appropriate labeling protocols will be applied, as described below for specific embodiments. Standard labeling protocols for nucleic acids are described, e.g., in Sambrook et al.; Kambara, H. et al. (1988) *BioTechnology* 6:816-821; Smith, L. et al. (1985) *Nuc. Acids Res.* 13:2399-2412; for polypeptides, see, e.g., Allen G. (1989) *Sequencing of Proteins and Peptides*, Elsevier, New York, especially chapter 5, and Greenstein and Winitz (1961) *Chemistry of the Amino Acids*, Wiley and Sons, New York. Carbohydrate labeling is described, e.g., in Chaplin and Kennedy (1986) *Carbohydrate Analysis: A Practical Approach*, IRL Press, Oxford. Labeling of other polymers will be performed by methods applicable to them as recognized by a person having ordinary skill in manipulating the corresponding polymer.

IAFP00000678

US 6,355,432 B1

23

In some embodiments, the target need not actually be labeled if a means for detecting where interaction takes place is available. As described below, for a nucleic acid embodiment, such may be provided by an intercalating dye which intercalates only into double stranded segments, e.g., where interaction occurs. See, e.g., Sheldon et al. U.S. Pat. No. 4,582,789.

In many uses, the target sequence will be absolutely homogeneous, both with respect to the total sequence and with respect to the ends of each molecule. Homogeneity with respect to sequence is important to avoid ambiguity. It is preferable that the target sequences of interest not be contaminated with a significant amount of labeled contaminating sequences. The extent of allowable contamination will depend on the sensitivity of the detection system and the inherent signal to noise of the system. Homogeneous contamination sequences will be particularly disruptive of the sequencing procedure.

However, although the target polynucleotide must have a unique sequence, the target molecules need not have identical ends. In fact, the homogeneous target molecule preparation may be randomly sheared to increase the numerical number of molecules. Since the total information content remains the same, the shearing results only in a higher number of distinct sequences which may be labeled and bind to the probe. This fragmentation may give a vastly superior signal relative to a preparation of the target molecules having homogeneous ends.

The signal for the hybridization is likely to be dependent on the numerical frequency of the target-probe interactions. If a sequence is individually found on a larger number of separate molecules a better signal will result. In fact, shearing a homogeneous preparation of the target may often be preferred before the labeling procedure is performed, thereby producing a large number of labeling groups associated with each subsequence.

C. Hybridization Conditions

The hybridization conditions between probe and target should be selected such that the specific recognition interaction, i.e., hybridization, of the two molecules is both sufficiently specific and sufficiently stable. See, e.g., Hames and Higgins (1985) *Nucleic Acid Hybridisation: A Practical Approach*, IRL Press, Oxford. These conditions will be dependent both on the specific sequence and often on the guanine and cytosine (GC) content of the complementary hybrid strands. The conditions may often be selected to be universally equally stable independent of the specific sequences involved. This typically will make use of a reagent such as an alkylammonium buffer. See, Wood et al. (1985) "Base Composition-independent Hybridization in Tetramethylammonium Chloride: A Method for Oligonucleotide Screening of Highly Complex Gene Libraries," *Proc. Natl. Acad. Sci. USA*, 82:1585-1588; and Krupov et al. (1989) "An Oligonucleotide Hybridization Approach to DNA Sequencing," *FEBS Letters*, 256:118-122; each of which is hereby incorporated herein by reference. An alkylammonium buffer tends to minimize differences in hybridization rate and stability due to GC content. By virtue of the fact that sequences then hybridize with approximately equal affinity and stability, there is relatively little bias in strength or kinetics of binding for particular sequences. Temperature and salt conditions along with other buffer parameters should be selected such that the kinetics of renaturation should be essentially independent of the specific target subsequence or oligonucleotide probe involved. In order to ensure this, the hybridization reactions will usually be performed in a single incubation of all the substrate matrices

24

together exposed to the identical same target probe solution under the same conditions.

Alternatively, various substrates may be individually treated differently. Different substrates may be produced, each having reagents which bind to target subsequences with substantially identical stabilities and kinetics of hybridization. For example, all of the high GC content probes could be synthesized on a single substrate which is treated accordingly. In this embodiment, the arylammonium buffers could be unnecessary. Each substrate is then treated in a manner such that the collection of substrates show essentially uniform binding and the hybridization data of target binding to the individual substrate matrix is combined with the data from other substrates to derive the necessary subsequence binding information. The hybridization conditions will usually be selected to be sufficiently specific such that the fidelity of base matching will be properly discriminated. Of course, control hybridizations should be included to determine the stringency and kinetics of hybridization.

D. Detection; VLSIPS™ Technology Scanning

The next step of the sequencing process by hybridization involves labeling of target polynucleotide molecules. A quickly and easily detectable signal is preferred. The VLSIPS™ Technology apparatus is designed to easily detect a fluorescent label, so fluorescent tagging of the target sequence is preferred. Other suitable labels include heavy metal labels, magnetic probes, chromogenic labels (e.g., phosphorescent labels, dyes, and fluorophores) spectroscopic labels, enzyme linked labels, radioactive labels, and labeled binding proteins. Additional labels are described in U.S. Pat. No. 4,366,241, which is incorporated herein by reference.

The detection methods used to determine where hybridization has taken place will typically depend upon the label selected above. Thus, for a fluorescent label a fluorescent detection step will typically be used. Pirrung et al. (1992) U.S. Pat. No. 5,143,854 and Ser. No. 07/624,120, now abandoned, describe apparatus and mechanisms for scanning a substrate matrix using fluorescence detection, but a similar apparatus is adaptable for other optically detectable labels.

The detection method provides a positional localization of the region where hybridization has taken place. However, the position is correlated with the specific sequence of the probe since the probe has specifically been attached or synthesized at a defined substrate matrix position. Having collected all of the data indicating the subsequences present in the target sequence, this data may be aligned by overlap to reconstruct the entire sequence of the target, as illustrated above.

It is also possible to dispense with actual labeling if some means for detecting the positions of interaction between the sequence specific reagent and the target molecule are available. This may take the form of an additional reagent which can indicate the sites either of interaction, or the sites of lack of interaction, e.g., a negative label. For the nucleic acid embodiments, locations of double strand interaction may be detected by the incorporation of intercalating dyes, or other reagents such as antibody or other reagents that recognize helix formation, see, e.g., Sheldon, et al. (1986) U.S. Pat. No. 4,582,789, which is hereby incorporated herein by reference.

E. Analysis

Although the reconstruction can be performed manually as illustrated above, a computer program will typically be used to perform the overlap analysis. A program may be written and run on any of a large number of different

US 6,355,432 B1

25

computer hardware systems. The variety of operating systems and languages useable will be recognized by a computer software engineer. Various different languages may be used, e.g., BASIC; C; PASCAL; etc. A simple flow chart of data analysis is illustrated in FIG. 1.

F. Substrate Reuse

Finally, after a particular sequence has been hybridized and the pattern of hybridization analyzed, the matrix substrate should be reusable and readily prepared for exposure to a second or subsequent target polynucleotides. In order to do so, the hybrid duplexes are disrupted and the matrix treated in a way which removes all traces of the original target. The matrix may be treated with various detergents or solvents to which the substrate, the oligonucleotide probes, and the linkages to the substrate are inert. This treatment may include an elevated temperature treatment, treatment with organic or inorganic solvents, modifications in pH, and other means for disrupting specific interaction. Thereafter, a second target may actually be applied to the recycled matrix and analyzed as before.

G. Non-Polynucleotide Aspects

Although the sequencing, fingerprinting, and mapping functions will make use of the natural sequence recognition property of complementary nucleotide sequences, the non-polynucleotide sequences typically require other sequence recognition reagents. These reagents will take the form, typically, of proteins exhibiting binding specificity, e.g., enzyme binding sites or antibody binding sites.

Enzyme binding sites may be derived from promoter proteins, restriction enzymes, and the like. See, e.g., Stryer, L. (1988) *Biochemistry*, W.H. Freeman, Palo Alto. Antibodies will typically be produced using standard procedures, see, e.g., Harlow and Lane (1988) *Antibodies: A Laboratory Manual*, Cold Spring Harbor Press, New York; and Goding (1986) *Monoclonal Antibodies: Principles and Practice*, (2d Ed.) Academic Press, San Diego.

Typically, an antigen, or collection of antigens are presented to an immune system. This may take the form of synthesized short polymers produced by the VLSIPS technology, or by the other synthetic means, or from isolation of natural products. For example, antigen for the polypeptides may be made by the VLSIPS technology, by standard peptide synthesis, by isolation of natural proteins with or without degradation to shorter segments, or by expression of a collection of short nucleic acids of random or defined sequences. See, e.g., Tuerk and Gold (1990) *Science* 249:505-510, for generation of a collection of randomly mutagenized oligonucleotides useful for expression.

The antigen or collection is presented to an appropriate immune system, e.g., to a whole animal as in a standard immunization protocol, or to a collection of immune cells or equivalent. In particular, see Ward et al. (1989) *Nature* 341:544-546; and Huse et al. (1989) *Science* 246:1275-1281, each of which is hereby incorporated herein by reference.

A large diversity of antibodies will be generated, some of which have specificities for the desired sequences. Antibodies may be purified having the desired sequence specificities by isolating the cells producing them. For example, a VLSIPS substrate with the desired antigens synthesized thereon may be used to isolate cells with cell surface reagents which recognize the antigens. The VLSIPS substrate may be used as an affinity reagent to select and recover the appropriate cells. Antibodies from those cells may be attached to a substrate using the caged biotin methodology, or by attaching a targeting molecule, e.g., an oligonucle-

26

otide. Alternatively, the supernatants from antibody producing cells can be easily assayed using a VLSIPS substrate to identify the cells producing the appropriate antibodies.

Although cells may be isolated, specific antibody molecules which perform the sequence recognition will also be sufficient. Preferably populations of antibody with a known specificity can be isolated. Supernatants from a large population of producing cells may be passed over a VLSIPS substrate to bind to the desired antigens attached to the substrate. When a sufficient density of antibody molecules are attached, they may be removed by an automated process, preferably as antibody populations exhibiting specificity of binding.

In one particular embodiment, a VLSIPS substrate, e.g., with a large plurality of fingerprint antigens attached thereto, is used to isolate antibodies from a supernatant of a population of cells producing antibodies to the antigens. Using the substrate as an affinity reagent, the antibodies will attach to the appropriate positionally defined antigens. The antibodies may be carefully removed therefrom, preferably by an automated system which retains their homogeneous specificities. The isolated antibodies can be attached to a new substrate in a positionally defined matrix pattern.

In a further embodiment, these spatially separated antibodies may be isolated using a specific targeting method for isolation. In this embodiment, a linker molecule which attaches to a particular portion of the antibody, preferably away from the binding site, can be attached to the antibodies. Various reagents will be used, including staphylococcus protein A or antibodies which bind to domains remote from the binding site. Alternatively, the antibodies in the population, before affinity purification, may be derivatized with an appropriate reagent compatible with new VLSIPS synthesis. A preferred reagent is a nucleotide which can serve as a linker to synthetic VLSIPS steps for synthesizing a specific sequence thereon. Then, by successive VLSIPS cycles, each of the antibodies attached to the defined antigen regions can have a defined oligonucleotide synthesized thereon and corresponding in area to the region of the substrate having each antigen attached. These defined oligonucleotides will be useful as targeting reagents to attach those antibodies possessing the same target sequence specificity at defined positions on a new substrate, by virtue of having bound to the antigen region, to a new VLSIPS substrate having the complementary target oligonucleotides positionally located on it. In this fashion, a VLSIPS substrate having the desired antigens attached thereto can be used to generate a second VLSIPS substrate with positionally defined reagents which recognize those antigens.

The selected antigens will typically be selected to be those which define particular functionalities or properties, so as to be useful for fingerprinting and other uses. They will also be useful for mapping and sequencing embodiments.

IV. FINGERPRINTING

A. General

Many of the procedures and techniques used in the polynucleotide sequencing section are also appropriate for fingerprinting applications. See, e.g., Pousika, et al. (1986) *Cold Spring Harbor Symposia on Quant. Biol.*, vol. LI, 131-139, Cold Spring Harbor Press, New York; which is hereby incorporated herein by reference. The fingerprinting method provided herein is based, in part, upon the ability to positionally localize a large number of different specific probes onto a single substrate. This high density matrix pattern provides the ability to screen for, or detect, a very large number of different sequences simultaneously. In fact, depending upon the hybridization conditions, fingerprinting

US 6,355,432 B1

27

to the resolution of virtually absolute matching of sequence is possible thereby approaching an absolute sequencing embodiment. And the sequencing embodiment is very useful in identifying the probes useful in further fingerprinting uses. For example, characteristic features of genetic sequences will be identified as being diagnostic of the entire sequence. However, in most embodiments, longer probe and target will be used, and for which slight mismatching may not need to be resolved.

B. Preparation of Substrate Matrix

A collection of specific probes may be produced by either of the methods described above in the section on sequencing. Specific oligonucleotide probes of desired lengths may be individually synthesized on a standard oligonucleotide synthesizer. The length of these probes is limited only by the ability of the synthesizer to continue to accurately synthesize a molecule. Oligonucleotides or sequence fragments may also be isolated from natural sources. Biological amplification methods may be coupled with synthetic synthesizing procedures such as, e.g., polymerase chain reaction.

In one embodiment, the individually isolated probes may be attached to the matrix at defined positions. These probe reagents may be attached by an automated process making use of the caged biotin methodology described in Ser. No. 07/612,671, or using photochemical reagents, see, e.g., Dattagupta et al. (1985) U.S. Pat. No. 4,542,102 and (1987) U.S. Pat. No. 4,713,326. Each individually purified reagent can be attached individually at specific locations on a substrate.

In another embodiment, the VLSIPS synthesizing technique may be used to synthesize the desired probes at specific positions on a substrate. The probes may be synthesized by successively adding appropriate monomer subunits, e.g., nucleotides, to generate the desired sequences.

In another embodiment, a relatively short specific oligonucleotide is used which serves as a targeting reagent for positionally directing the sequence recognition reagent. For example, the sequence specific reagents having a separate additional sequence recognition segment (usually of a different polymer from the target sequence) can be directed to target oligonucleotides attached to the substrate. By use of non-natural targeting reagents, e.g., unusual nucleotide analogues which pair with other unnatural nucleotide analogues and which do not interfere with natural nucleotide interactions, the natural and non-natural portions can coexist on the same molecule without interfering with their individual functionalities. This can combine both a synthetic and biological production system analogous to the technique for targeting monoclonal antibodies to locations on a VLSIPS substrate at defined positions. Unnatural optical isomers of nucleotides may be useful unnatural reagents subject to similar chemistry, but incapable of interfering with the natural biological polymers. See also, Ser. No. 07/626,730, which is hereby incorporated herein by reference.

After the separate substrate attached reagents are attached to the targeting segment, the two are crosslinked, thereby permanently attaching them to the substrate. Suitable crosslinking reagents are known, see, e.g., Dattagupta et al. (1985) U.S. Pat. No. 4,542,102 and (1987) "Coupling of nucleic acids to solid support by photochemical methods," U.S. Pat. No. 4,713,326, each of which is hereby incorporated herein by reference. Similar linkages for attachment of proteins to a solid substrate are provided, e.g., in Merrifield (1986) *Science* 232:341-347, which is hereby incorporated herein by reference.

28

C. Labeling Target Nucleotides

The labeling procedures used in the sequencing embodiments will also be applicable in the fingerprinting embodiments. However, since the fingerprinting embodiments often will involve relatively large target molecules and relatively short oligonucleotide probes, the amount of signal necessary to incorporate into the target sequence may be less critical than in the sequencing applications. For example, a relatively long target with a relatively small number of labels per molecule may be easily amplified or detected because of the relatively large target molecule size.

In various embodiments, it may be desired to cleave the target into smaller segments as in the sequencing embodiments. The labeling procedures and cleavage techniques described in the sequencing embodiments would usually also be applicable here.

D. Hybridization Conditions

The hybridization conditions used in fingerprinting embodiments will typically be less critical than for the sequencing embodiments. The reason is that the amount of mismatching which may be useful in providing the fingerprinting information would typically be far greater than that necessary in sequencing uses. For example, Southern hybridizations do not typically distinguish between slightly mismatched sequences. Under these circumstances, important and valuable information may be arrived at with less stringent hybridization conditions while providing valuable fingerprinting information. However, since the entire substrate is typically exposed to the target molecule at one time, the binding affinity of the probes should usually be of approximately comparable levels. For this reason, if oligonucleotide probes are being used, their lengths should be approximately comparable and will be selected to hybridize under conditions which are common for most of the probes on the substrate. Much as in a Southern hybridization, the target and oligonucleotide probes are of lengths typically greater than about 25 nucleotides. Under appropriate hybridization conditions, e.g., typically higher salt and lower temperature, the probes will hybridize irrespective of imperfect complementarity. In fact, with probes of greater than, e.g., about fifty nucleotides, the difference in stability of different sized probes will be relatively minor.

Typically the fingerprinting is merely for probing similarity or homology. Thus, the stringency of hybridization can usually be decreased to fairly low levels. See, e.g., Wetmur and Davidson (1968) "Kinetics of Renaturation of DNA," *J. Mol. Biol.*, 31:349-370; and Kanehisa, M. (1984) *Nuc. Acids Res.*, 12:203-213.

E. Detection; VLSIPS™ Technology Scanning

Detection methods will be selected which are appropriate for the selected label. The scanning device need not necessarily be digitized or placed into a specific digital database, though such would most likely be done. For example, the analysis in fingerprinting could be photographic. Where a standardized fingerprint substrate matrix is used, the pattern of hybridizations may be spatially unique and may be compared photographically. In this manner, each sample may have a characteristic pattern of interactions and the likelihood of identical patterns will preferably be such low frequency that the fingerprint pattern indeed becomes a characteristic pattern virtually as unique as an individual's fingertip fingerprint. With a standardized substrate, every individual could be, in theory, uniquely identifiable on the basis of the pattern of hybridizing to the substrate.

Of course, the VLSIPS™ Technology scanning apparatus may also be useful to generate a digitized version of the fingerprint pattern. In this way, the identification pattern can

US 6,355,432 B1

29

be provided in a linear string of digits. This sequence could also be used for a standardized identification system providing significant useful medical transferability of specific data. In one embodiment, the probes used are selected to be of sufficiently high resolution to measure the antigens of the major histocompatibility complex. It might even be possible to provide transplantation matching data in a linear stream of data. The fingerprinting data may provide a condensed version, or summary, of the linear genetic data, or any other information data base.

F. Analysis

The analysis of the fingerprint will often be much simpler than a total sequence determination. However, there may be particular types of analysis which will be substantially simplified by a selected group of probes. For example, probes which exhibit particular populational heterogeneity may be selected. In this way, analysis may be simplified and practical utility enhanced merely by careful selection of the specific probes and a careful matrix layout of those probes.

G. Substrate Reuse

As with the sequencing application, the fingerprinting usages may also take advantage of the reusability of the substrate. In this way, the interactions can be disrupted, the substrate treated, and the renewed substrate is equivalent to an unused substrate.

H. Non-polynucleotide Aspects

Besides polynucleotide applications, the fingerprinting analysis may be applied to other polymers, especially polypeptides, carbohydrates, and other polymers, both organic and inorganic. Besides using the fingerprinting method for analyzing a particular polymer, the fingerprinting method may be used to characterize various samples. For example, a cell or population of cells may be tested for their expression of specific antigens or their mRNA sequence intent. For example, a T-cell may be classified by virtue of its combination of expressed surface antigens. With specific reagents which interact with these antigens, a cell or a population of cells or a lysed cell may be exposed to a VLSIPS substrate. The biological sample may be classified or characterized by analyzing the pattern of specific interaction.

This may be applicable to a cell or tissue type, to the messenger RNA population expressed by a cell to the genetic content of a cell, or to virtually any sample which can be classified and/or identified by its combination of specific molecular properties.

The ability to generate a high density means for screening the presence or absence of specific interactions allows for the possibility of screening for, if not saturating, all of a very large number of possible interactions. This is very powerful in providing the means for testing the combinations of molecular properties which can define a class of samples. For example, a species of organism may be characterized by its DNA sequences, e.g., a genetic fingerprint. By using a fingerprinting method, it may be determined that all members of that species are sufficiently similar in specific sequences that they can be easily identified as being within a particular group. Thus, newly defined classes may be resolved by their similarity in fingerprint patterns. Alternatively, a non-member of that group will fail to share those many identifying characteristics. However, since the technology allows testing of a very large number of specific interactions, it also provides the ability to more finely distinguish between closely related different cells or samples. This will have important applications in diagnosing viral, bacterial, and other pathological on nonpathological infections.

30

In particular, cell classification may be defined by any of a number of different properties. For example, a cell class may be defined by its DNA sequences contained therein. This allows species identification for parasitic or other infections. For example, the human cell is presumably genetically distinguishable from a monkey cell, but different human cells will share many genetic markers. At higher resolution, each individual human genome will exhibit unique sequences that can define it as a single individual.

Likewise, a developmental stage of a cell type may be definable by its pattern of expression of messenger RNA. For example, in particular stages of cells, high levels of ribosomal RNA are found whereas relatively low levels of other types of messenger RNAs may be found. The high resolution distinguishability provided by this fingerprinting method allows the distinction between cells which have relatively minor differences in its expressed mRNA population. Where a pattern is shown to be characteristic of a stage, a stage may be defined by that particular pattern of messenger RNA expression.

In a similar manner, the antigenic determinants found on a protein may very well define the cell class. For example, immunological T-cells are distinguishable from B-cells because, in part, the cell surface antigens on the cell types are distinguishable. Different T-cell subclasses can be also distinguished from one another by whether they contain particular T-cell antigens. The present invention provides the possibility for high resolution testing of many different interactions simultaneously, and the definition of new cell types will be possible.

The high resolution VLSIPS™ substrate may also be used as a very powerful diagnostic tool to test the combination of presence, of a plurality of different assays from a biological sample. For example, a cancerous condition may be indicated by a combination of various different properties found in the blood. For example, a cancerous condition may be indicated by a combination of expression of various soluble antigens found in the blood along with a high number of various cellular antigens found on lymphocytes and/or particular cell degradation products. With a substrate as provided herein, a large number of different features can be simultaneously performed on a biological sample. In fact, the high resolution of the test will allow more complete characterization of parameters which define particular diseases. Thus, the power of diagnostic tests may be limited by the extent of statistical correlation with a particular condition rather than with the number of antigens or interactions which are tested. The present invention provides the means to generate this large universe of possible reagents and the ability to actually accumulate that correlative data.

In another embodiment, a substrate as provided herein may be used for genetic screening. This would allow for simultaneous screening of thousands of genetic markers. As the density of the matrix is increased, many more molecules can be simultaneously tested. Genetic screening then becomes a simpler method as the present invention provides the ability to screen for thousands, tens of thousands, and hundreds of thousands, even millions of different possible genetic features. However, the number of high correlation genetic markers for conditions numbers only in the hundreds. Again, the possibility for screening a large number of sequences provides the opportunity for generating the data which can provide correlation between sequences and specific conditions or susceptibility. The present invention provides the means to generate extremely valuable correlations useful for the genetic detection of the causative mutation leading to medical conditions. In still another embodiment, the present invention would be applicable to distinguishing two individuals having identical genetic compositions. The antibody population within an individual is dependent both on genetic and historical factors. Each

IAFP00000682

US 6,355,432 B1

31

individual experiences a unique exposure to various infectious agents, and the combined antibody expression is partly determined thereby. Thus, individuals may also be fingerprinted by their immunological content, either of actively expressed antibodies, or their immunological memory. Similar sorts of immunological and environmental histories may be useful for fingerprinting, perhaps in combination with other screening properties. In particular, the present invention may be useful for screening allergic reactions or susceptibilities, and a simple IgE specificity test may be useful in determining a spectrum of allergies.

With the definition of new classes of cells, a cell sorter will be used to purify them. Moreover, new markers for defining that class of cells will be identified. For example, where the class is defined by its RNA content, cells may be screened by antisense probes which detect the presence or absence of specific sequences therein. Alternatively, cell lysates may provide information useful in correlating intracellular properties with extracellular markers which indicate functional differences. Using standard cell sorter technology with a fluorescence or labeled antisense probe which recognizes the internal presence of the specific sequences of interest, the cell sorter will be able to isolate a relatively homogeneous population of cells possessing the particular marker. Using successive probes the sorting process should be able to select for cells having a combination of a large number of different markers.

In a non-polynucleotide embodiment, cells may be defined by the presence of other markers. The markers may be carbohydrates, proteins, or other molecules. Thus, a substrate having particular specific reagents, e.g., antibodies, attached to it should be able to identify cells having particular patterns of marker expression. Of course, combinations of these made be utilized and a cell class may be defined by a combination of its expressed mRNA, its carbohydrate expression, its antigens, and other properties. This fingerprinting should be useful in determining the physiological state of a cell or population of cells.

Having defined a cell type whose function or properties are defined by the reagents attachable to a VLSIPS substrate, such as cellular antigens, these structural manifestations of function may be used to sort cells to generate a relatively homogeneous population of that class of cells. Standard cell sorter technology may be applied to purify such a population, see, e.g., Dangi, J. and Herzenberg (1982) "Selection of hybridomas and hybridoma variants using the fluorescence activated cell sorter," *J. Immunological Methods* 52:1-14; and Becton Dickinson, Fluorescence Activated Cell Sorter Division, San Jose, Calif., and Coulter Diagnostics, Hialeah, Fla.

With the fingerprinting method an identification means arises from mosaicism problems in an organism. A mosaic organism is one whose genetic content in different cells is significantly different. Various clonal populations should have similar genetic fingerprints, though different clonal populations may have different genetic contents. See, for example, Suzuki et al. *An Introduction to Genetic Analysis* (4th Ed.), Freeman and Co., New York, which is hereby incorporated herein by reference. However, this problem should be a relatively rare problem and could be more carefully evaluated with greater experience using the fingerprinting methods.

The invention will also find use in detecting changes, both genetic and antigenic, e.g., in a rapidly "evolving" protozoa infection, or similarly changing organism.

V. MAPPING

A. General

The use of the present invention for mapping parallels its use for fingerprinting and sequencing. Where a polymer is a linear molecule, the mapping provides the ability to locate

32

particular segments along the length of the polymer. Branched polymers can be treated as a series of individual linear polymers. The mapping provides the ability to locate, in a relative sense, the order of various subsequences. This may be achieved using at least two different approaches.

The first approach is to take the large sequence and fragment it at specific points. The fragments are then ordered and attached to a solid substrate. For example, the clones resulting from a chromosome walking process may be individually attached to the substrate by methods, e.g., caged biotin techniques, indicated earlier. Segments of unknown map position will be exposed to the substrate and will hybridize to the segment which contains that particular sequence. This procedure allows the rapid determination of a number of different labeled segments, each mapping requiring only a single hybridization step once the substrate is generated. The substrate may be regenerated by removal of the interaction, and the next mapping segment applied.

In an alternative method, a plurality of subsequences can be attached to a substrate. Various short probes may be applied to determine which segments may contain particular overlaps. The theoretical basis and a description of this mapping procedure is contained in, e.g., Evans et al. 1989 "Physical Mapping of Complex Genomes by Cosmid Multiplex Analysis," *Proc. Natl. Acad. Sci. USA* 86:5030-5034, and other references cited above in the Section labeled "Overall Description." Using this approach, the details of the mapping embodiment are very similar to those used in the fingerprinting embodiment.

B. Preparation of Substrate Matrix

The substrate may be generated in either of the methods generally applicable in the sequencing and fingerprinting embodiments. The substrate may be made either synthetically, or by attaching otherwise purified probes or sequences to the matrix. The probes or sequences may be derived either from synthetic or biological means. As indicated above, the solid phase substrate synthetic methods may be utilized to generate a matrix with positionally defined sequences. In the mapping embodiment, the importance of saturation of all possible subsequences of a preselected length is far less important than in the sequencing embodiment, but the length of the probes used may be desired to be much longer. The processes for making a substrate which has longer oligonucleotide probes should not be significantly different from those described for the sequencing embodiments, but the optimization parameters may be modified to comply with the mapping needs.

C. Labeling

The labeling methods will be similar to those applicable in sequencing and fingerprinting embodiments. Again, it may be desirable to fragment the target sequences.

D. Hybridization/Specific Interaction

The specificity of interaction between the targets and probe would typically be closer to those used for fingerprinting embodiments, where homology is more important than absolute distinguishability of high fidelity complementary hybridization. Usually, the hybridization conditions will be such that merely homologous segments will interact and provide a positive signal. Much like the fingerprinting embodiment, it may be useful to measure the extent of homology by successive incubations at higher stringency conditions. Or, a plurality of different probes, each having various levels of homology may be used. In either way, the spectrum of homologies can be measured.

Where non-nucleic acid hybridization is involved, the specific interactions may also be compared in a fingerprint-like manner. The specific reagents may have less specificity, e.g., monoclonal antibodies which recognize a broader spectrum of sequences may be utilized relative to a sequencing embodiment. Again, the specificity of interaction may be measured under various conditions of increasing stringency

US 6,355,432 B1

33

to determine the spectrum of matching across the specific probes selected, or a number of different stringency reagents may be included to indicate the binding affinity.

E. Detection

The detection methods used in the mapping procedure will be virtually identical to those used in the fingerprinting embodiment. The detection methods will be selected in combination with the labeling methods.

F. Analysis

The analysis of the data in a mapping embodiment will typically be somewhat different from that in fingerprinting.

The fingerprinting embodiment will test for the presence or absence of specific or homologous segments. However, in the mapping embodiment, the existence of an interaction is coupled with some indication of the location of the interaction. The interaction is mapped in some manner to the physical polymer sequence. Some means for determining the relative positions of different probes is performed. This may be achieved by synthesis of the substrate in pattern, or may result from analysis of sequences after they have been attached to the substrate.

For example, the probes may be randomly positioned at various locations on the substrate. However, the relative positions of the various reagents in the original polymer may be determined by using short fragments, e.g., individually, as target molecules which determine the proximity of different probes. By an automated system of testing each different short fragment of the original polymer, coupled with proper analysis, it will be possible to determine which probes are adjacent one another on the original target sequence and correlate that with positions on the matrix. In this way, the matrix is useful for determining the relative locations of various new segments in the original target molecule. This sort of analysis is described in Evans, and the related references described above.

G. Substrate Reuse

The substrate should be reusable in the manner described in the fingerprinting section. The substrate is renewed by removal of the specific interactions and is washed and prepared for successive cycles of exposure to new target sequences.

H. Non-polynucleotide Aspects

The mapping procedure may be used on other molecules than polynucleotides. Although hybridization is one type of specific interaction which is clearly useful for use in this mapping embodiment, antibody reagents may also be very useful.

In the same way that polypeptide sequencing or other polymers may be sequenced by the reagents and techniques described in the sequencing section and fingerprinting section, the mapping embodiment may also be used similarly.

In another form of mapping, as described above in the fingerprinting section, the developmental map of a cell or biological system may be measured using fingerprinting type technology. Thus, the mapping may be along a temporal dimension rather than along a polymer dimension. The mapping or fingerprinting embodiments may also be used in determining the genetic rearrangements which may be genetically important, as in lymphocyte and B-cell development. In another example, various rearrangements or chromosomal dislocations may be tested by either the fingerprinting or mapping methods. These techniques are similar in many respects and the fingerprinting and mapping embodiments may overlap in many respects.

VI. ADDITIONAL SCREENING AND APPLICATIONS

A. Specific Interactions

As originally indicated in the parent filing of VLSIPS™ Technology, the production of a high density plurality of

34

spatially segregated polymers provides the ability to generate a very large universe or repertoire of individually, and distinct sequence possibilities. As indicated above, particular oligonucleotides may be synthesized in automated fashion at specific locations on a matrix. In fact, these oligonucleotides may be used to direct other molecules to specific locations by linking specific oligonucleotides to other reagents which are in batch exposed to the matrix and hybridized in a complementary fashion to only those locations where the complementary oligonucleotide has been synthesized on the matrix. This allows for spatially attaching a plurality of different reagents onto the matrix instead of individually attaching each separate reagent at each specific location. Although the caged biotin method allows automated attachment, the speed of the caged biotin attachment process is relatively slow and requires a separate reaction for each reagent being attached. By use of the oligonucleotide method, the specificity of position can be done in an automated and parallel fashion. As each reagent is produced, instead of directly attaching each reagent at each desired position, the reagent may be attached to a specific desired complementary oligonucleotide which will ultimately be specifically directed toward locations on the matrix having a complementary oligonucleotide attached thereat.

In addition, the technology allows screening for specificity of interaction with particular reagents. For example, the oligonucleotide sequence specificity of binding of a potential reagent may be tested by presenting to the reagent all of the possible subsequences available for binding.

Although secondary or higher order sequence specific features might not be easily screenable using this technology, it does provide a convenient, simple, quick, and thorough screen of interactions between a reagent and its target recognition sequences. See, e.g., Pfeifer et al. (1989) *Science* 246:810-812.

For example, the interaction of a promoter protein with its target binding sequence may be tested for many different, or all, possible binding sequences. By testing the strength of interactions under various different conditions, the interaction of the promoter protein with each of the different potential binding sites may be analyzed. The spectrum of strength of interactions with each different potential binding site may provide significant insight into the types of features which are important in determining specificity.

An additional example of a sequence specific interaction between reagents is the testing of binding of a double stranded nucleic acid structure with a single stranded oligonucleotide. Often, a triple stranded structure is produced which has significant aspects of sequence specificity. Testing of such interactions with either sequences comprising only natural nucleotides, or perhaps the testing of nucleotide analogs may be very important in screening for particularly useful diagnostic or therapeutic reagents. See, e.g., Häner and Dervan (1990) *Biochemistry* 29:9761-6765, and references therein.

B. Sequence Comparisons

Once a gene is sequenced, the present invention provides a means to compare alleles or related sequences to locate and identify differences from the control sequence.

This would be extremely useful in further analysis of genetic variability at a specific gene locus.

C. Categorizations

As indicated above in the fingerprinting and mapping embodiments, the present invention is also useful in defining specific stages in the temporal sequence of cells, e.g., development, and the resulting tissues within an organism. For example, the developmental stage of a cell, or population of cells, can be dependent upon the expression of particular messenger RNAs or cellular antigens. The screening procedures provided allow for high resolution definition of new classes of cells. In addition, the temporal develop-

US 6,355,432 B1

35

ment of particular cells will be characterized by the presence or expression of various mRNAs. Means to simultaneously screen a plurality or very large number of different sequences are provided. The combination of different markers made available dramatically increases the ability to distinguish fairly closely related cell types. Other markers may be combined with markers and methods made available herein to define new classifications of biological samples, e.g., based upon new combinations of markers.

The presence or absence of particular marker sequences will be used to define temporal developmental stages.

Once the stages are defined, fairly simple methods can be applied to actually purify those particular cells. For example, antisense probes or recognition reagents may be used with a cell sorter to select those cells containing or expressing the critical markers. Alternatively, the expression of those sequences may result in specific antigens which may also be used in defining cell classes and sorting those cells away from others. In this way, for example, it should be possible to select a class of omnipotent immune system cells which are able to completely regenerate a human immune system.

Based upon the cellular classes defined by the parameters made available by this technology, purified classes of cells having identifiable differences, structural or functional, are made available.

In an alternative embodiment, a plurality of antigens or specific binding proteins attached to the substrate may be used to define particular cell types. For example, subclasses of T-cells are defined, in part, by the combination of expressed cell surface antigens. The present invention allows for the simultaneous screening of a large plurality of different antigens together. Thus, higher resolution classification of different T-cell subclasses becomes possible and, with the definitions and functional differences which correlate with those antigenic or other parameters, the ability to purify those cell types becomes available. This is applicable not only to T-cells, but also to lymphocyte cells, or even to freely circulating cells. Many of the cells for which this would be most useful will be immobile cells found in particular tissues or organs. Tumor cells will be diagnosed or detected using these fingerprinting techniques. Coupled with a temporal change in structure, developmental classes may also be selected and defined using these technologies. The present invention also provides the ability not only to define new classes of cells based upon functional or structural differences, but it also provides the ability to select or purify populations of cells which share these particular properties. Standard cell sorting procedures using antibody markers may be used to detect extracellular features. Intracellular features would also be detectable by introducing the label reagents into the cell. In particular, antisense DNA or RNA molecules may be introduced into a cell to detect RNA sequences therein. See, e.g., Weintraub (1990) *Scientific American* 262:40-46.

D. Statistical Correlations

In an additional embodiment, the present invention also allows for the high resolution correlation of medical conditions with various different markers. For example, the presently available technology, when applied to amniocentesis or other genetic screening methods, typically screens for tens of different markers at most. The present invention allows simultaneous screening for tens, hundreds, thousands, tens of thousands, hundreds of thousands, and even millions of different genetic sequences. Thus, applying the fingerprinting methods of the present invention to a sufficiently large population allows detailed statistical analysis to be made, thereby correlating particular medical conditions with particular markers, typically antigenic or genetic. Tumor specific antigens will be identified using the present invention.

Various medical conditions may be correlated against an enormous data base of the sequences within an individual.

36

Genetic propensities and correlations then become available and high resolution genetic predictability and correlation become much more easily performed. With the enormous data base, the reliability of the predictions is also better tested. Particular markers which are partially diagnostic of particular medical conditions or medical susceptibilities will be identified and provide direction in further studies and more careful analysis of the markers involved. Of course, as indicated above in the sequencing embodiment, the present invention will find much use in intense sequencing projects. For example, sequencing of the entire human genome in the human genome project will be greatly simplified and enabled by the present invention.

VI. FORMATION OF SUBSTRATE

The substrate is provided with a pattern of specific reagents which are positionally localized on the surface of the substrate. This matrix of positions is defined by the automated system which produces the substrate. The instrument will typically be one similar to that described in Pirrung et al. (1992) U.S. Pat. No. 5,143,854, and Ser. No. 07/624,120, now abandoned. The instrumentation described therein is directly applicable to the applications used here. In particular, the apparatus comprises a substrate, typically a silicon containing substrate, on which positions on the surface may be defined by a coordinate system of positions. These positions can be individually addressed or detected by the VLSIPS™ Technology apparatus.

Typically, the VLSIPS™ Technology apparatus uses optical methods used in semiconductor fabrication applications. In this way, masks may be used to photo-activate positions for attachment or synthesis of specific sequences on the substrate. These manipulations may be automated by the types of apparatus described in Pirrung et al. (1992) U.S. Pat. No. 5,143,854 and Ser. No. 07/624,120, now abandoned.

Selectively removable protecting groups allow creation of well defined areas of substrate surface having differing reactivities. Preferably, the protecting groups are selectively removed from the surface by applying a specific activator, such as electromagnetic radiation of a specific wavelength and intensity. More preferably, the specific activator exposes selected areas of surface to remove the protecting groups in the exposed areas.

Protecting groups of the present invention are used in conjunction with solid phase oligomer syntheses, such as peptide syntheses using natural or unnatural amino acids, nucleotide syntheses using deoxyribonucleic and ribonucleic acids, oligosaccharide syntheses, and the like. In addition to protecting the substrate surface from unwanted reaction, the protecting groups block a reactive end of the monomer to prevent self-polymerization. For instance, attachment of a protecting group to the amino terminus of an activated amino acid, such as the N-hydroxysuccinimide-activated ester of the amino acid prevents the amino terminus of one monomer from reacting with the activated ester portion of another during peptide synthesis.

Alternatively, the protecting group may be attached to the carboxyl group of an amino acid to prevent reaction at this site. Most protecting groups can be attached to either the amino or the carboxyl group of an amino acid, and the nature of the chemical synthesis will dictate which reactive group will require a protecting group. Analogously, attachment of a protecting group to the 5'-hydroxyl group of a nucleoside during synthesis using for example, phosphate-triester coupling chemistry, prevents the 5'-hydroxyl of one nucleoside from reacting with the 3'-activated phosphate-triester of another.

Regardless of the specific use, protecting groups are employed to protect a moiety on a molecule from reacting

US 6,355,432 B1

37

with another reagent. Protecting groups of the present invention have the following characteristics: they prevent selected reagents from modifying the group to which they are attached; they are stable (that is, they remain attached) to the synthesis reaction conditions; they are removable under conditions that do not adversely affect the remaining structure; and once removed, do not react appreciably with the surface or surface-bound oligomer. The selection of a suitable protecting group will depend, of course, on the chemical nature of the monomer unit and oligomer, as well as the specific reagents they are to protect against.

In a preferred embodiment, the protecting groups will be photoactivatable. The properties and uses of photoreactive protecting compounds have been reviewed. See, McCray et al., *Ann. Rev. of Biophys. and Biophys. Chem.* (1989) 18:239-270, which is incorporated herein by reference. Preferably, the photosensitive protecting groups will be removable by radiation in the ultraviolet (UV) or visible portion of the electromagnetic spectrum. More preferably, the protecting groups will be removable by radiation in the near UV or visible portion of the spectrum. In some embodiments, however, activation may be performed by other methods such as localized heating, electron beam lithography, laser pumping, oxidation or reduction with microelectrodes, and the like. Sulfonyl compounds are suitable reactive groups for electron beam lithography. Oxidative or reductive removal is accomplished by exposure of the protecting group to an electric current source, preferably using microelectrodes directed to the predefined regions of the surface which are desired for activation. A more detailed description of these protective groups is provided in Ser. No. 07/624,120, now abandoned, which is hereby incorporated herein by reference.

The density of reagents attached to a silicon substrate may be varied by standard procedures. The surface area for attachment of reagents may be increased by modifying the silicon surface. For example, a matte surface may be machined or etched on the substrate to provide more sites for attachment of the particular reagents. Another way to increase the density of reagent binding sites is to increase the derivitization density of the silicon. Standard procedures for achieving this are described, below.

One method to control the derivatization density is to highly derivatize the substrate with photochemical groups at high density. The substrate is then photolyzed for various predetermined times, which photoactivate the groups at a measurable rate, and react them with a capping reagent. By this method, the density of linker groups may be modulated by using a desired time and intensity of photoactivation.

In many applications, the number of different sequences which may be provided may be limited by the density and the size of the substrate on which the matrix pattern is generated. In situations where the density is insufficiently high to allow the screening of the desired number of sequences, multiple substrates may be used to increase the number of sequences tested. Thus, the number of sequences tested may be increased by using a plurality of different substrates. Because the VLSIPS apparatus is almost fully automated, increasing the number of substrates does not lead to a significant increase in the number of manipulations which must be performed by humans. This again leads to greater reproducibility and speed in the handling of these multiple substrates.

A. Instrumentation

The concept of using VLSIPS™ Technology generally allows a pattern or a matrix of reagents to be generated. The procedure for making the pattern is performed by any of a number of different methods. An apparatus and instrumentation useful for generating a high density VLSIPS substrate is described in detail in Pirrung et al. (1992) U.S. Pat. No. 5,143,854 and Ser. No. 07/624,120, now abandoned.

38

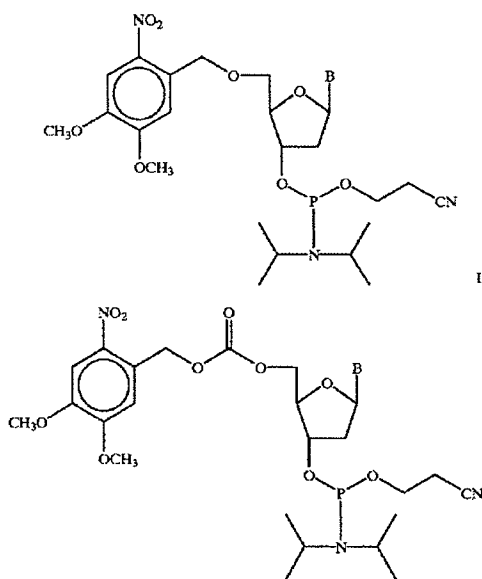
B. Binary Masking

The details of the binary masking are described in an accompanying application filed simultaneously with this, Ser. No. 07/624,120, now abandoned, whose specification is incorporated herein by reference.

For example, the binary masking technique allows for producing a plurality of sequences based on the selection of either of two possibilities at any particular location. By a series of binary masking steps, the binary decision may be the determination, on a particular synthetic cycle, whether or not to add any particular one of the possible subunits. By treating various regions of the matrix pattern in parallel, the binary masking strategy provides the ability to carry out spatially addressable parallel synthesis.

C. Synthetic Methods

The synthetic methods in making a substrate are described in the parent application, Pirrung et al. (1992) U.S. Pat. No. 5,143,854. The construction of the matrix pattern on the substrate will typically be generated by the use of photo-sensitive reagents. By use of photo-lithographic optical methods, particular segments of the substrate can be irradiated with light to activate or deactivate blocking agents, e.g., to protect or deprotect particular chemical groups. By an appropriate sequence of photo-exposure steps at appropriate times with appropriate masks and with appropriate reagents, the substrates can have known polymers synthesized at positionally defined regions on the substrate. Methods for synthesizing various substrates are described in Pirrung et al. (1992) U.S. Pat. No. 5,143,854 and Ser. No. 07/624,120, now abandoned. By a sequential series of these photo-exposure and reaction manipulations, a defined matrix pattern of known sequences may be generated, and is typically referred to as a VLSIPS™ Technology substrate. In the nucleic acid synthesis embodiment, nucleosides used in the synthesis of DNA by photolytic methods will typically be one of the two forms shown below:

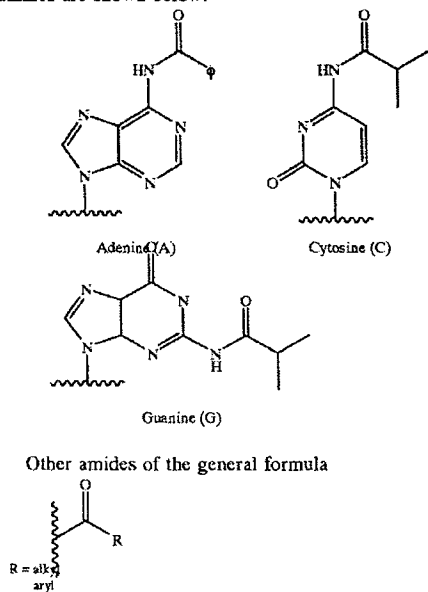


B=Adenine, Cytosine, Guanine, or Thymine

US 6,355,432 B1

39

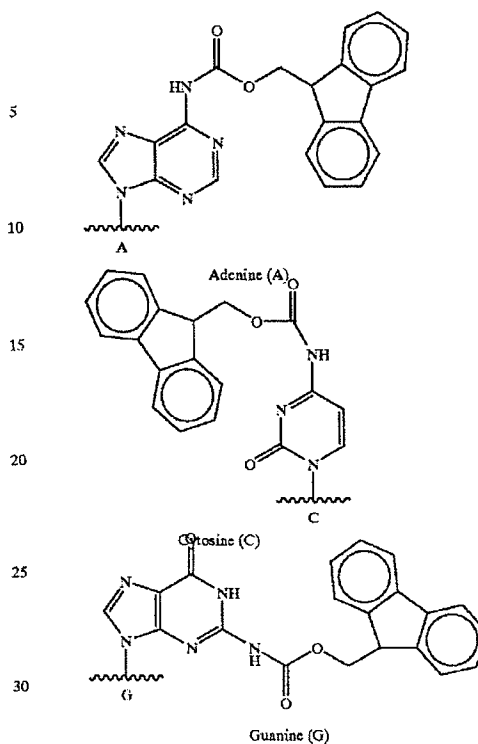
In I, the photolabile group at the 5' position is abbreviated NV (nitroveratryl) and in II, the group is abbreviated NVOC (nitroveratryl oxycarbonyl). Although not shown in FIG. C, the bases (adenine, cytosine, and guanine) contain exocyclic NH_2 groups which must be protected during DNA synthesis. 5
Thymine contains no exocyclic NH_2 and therefore requires no protection. The standard protecting groups for these amines are shown below:



where R may be alkyl or aryl have been used.

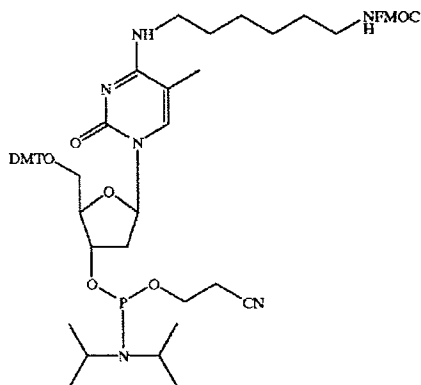
Another type of protecting group FMOC (9-fluorenyl methoxycarbonyl) is currently being used to protect the exocyclic amines of the three bases:

40



The advantage of the FMOC group is that it is removed 35 under mild conditions (dilute organic bases) and can be used for all three bases. The amide protecting groups require more harsh conditions to be removed (NH_3/MeOH with heat).

Nucleosides used as 5'-OH probes, useful in verifying correct VLSIPS synthetic function, include, for example, the following:



III

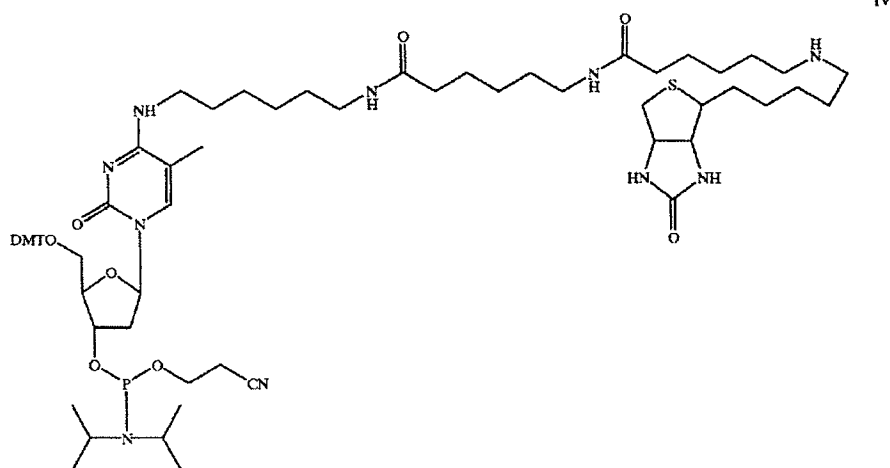
IAFP00000687

US 6,355,432 B1

41

42

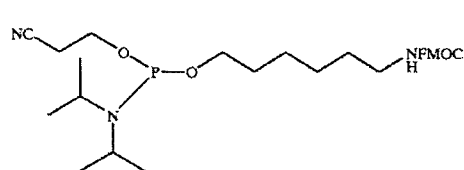
-continued



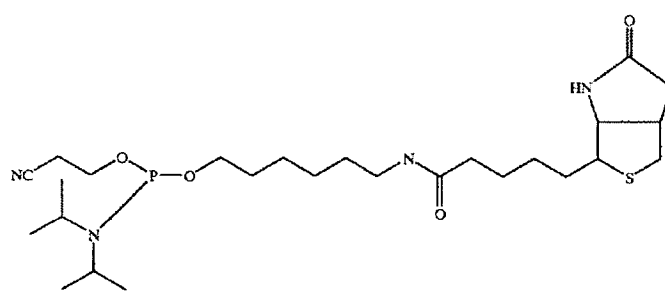
25

These compounds are used to detect where on a substrate photolysis has occurred by the attachment of either III or V to the newly generated 5'-OH. In the case of III, after the phosphate attachment is made, the substrate is treated with a dilute base to remove the Fmoc group. The resulting amine can be reacted with FITC and the substrate examined

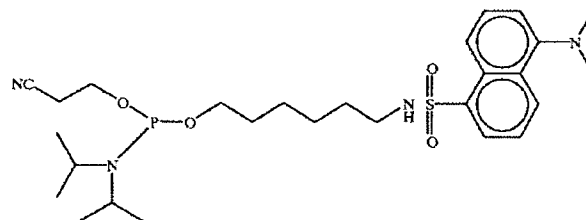
by fluorescence microscopy. This indicates the proper generation of a 5'-OH. In the case of compound IV, after the phosphate attachment is made, the substrate is treated with FITC labeled streptavidin and the substrate again may be examined by fluorescence microscopy. Other probes, although not nucleoside based, have included the following:



V



VI



VII

IAFP00000688